CALaMo:

A CONSTRUCTIONIST PERSPECTIVE ON THE ANALYSIS OF LINGUISTIC BEHAVIOUR OF LANGUAGE MODELS

Candidate:
LUDOVICA PANNITTO

Supervisor:
DR. AURELIE HERBELOT

Thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in the
DOCTORATE IN COGNITIVE AND BRAIN SCIENCES

34th cycle
CIMeC - Centre for Mind/Brain Sciences
University of Trento

Language must speak for itself.

— Joseph Kosuth, 1991

The word in language is half someone else's. It becomes one's "own" only when the speaker populates it with his own intentions, his own accent, when he appropriates the word, adapting it to his own semantic and expressive intention. Prior to this moment of appropriation, the word does not exist in a neutral and impersonal language... but rather it exists in other people's mouths, in other people's contexts, serving other people's intentions; it is from there that one must take the word, and make it one's own.

— Mikhail Bakhtin, *The Dialogic Imagination: Four Essay*, 1975

ABSTRACT

In recent years, Neural Language Models (NLMs) have consistently demonstrated increasing linguistic abilities. However, the extent to which such networks can actually *learn grammar* remains an object of investigation, and experimental results are often inconclusive.

Notably, the mainstream evaluation framework in which NLMs are tested seems largely based on Generative Grammar and nativist principles, and a shared constructionist approach on the matter has not yet emerged: this is at odds with the fact that usage-based theories are actually better suited to inspect the behaviour of such models.

The main contribution of this thesis is the introduction of CALaMo, a novel framework for evaluating Neural Language Models' linguistic abilities, using a constructionist approach. We especially aim at formalizing the relationship between the computational modelling phase and the underlying linguistic theory, thus allowing a more refined and informed discussion of settings and results.

We focus on two specific areas that, we believe, are currently not easily tractable within the mainstream evaluation framework.

The first scenario deals with language acquisition from child-directed data. Our main experimental result shows how it is possible to follow schematization paths during the acquisition process of the model, and how this relates to core hypotheses in constructionist theories.

The second scenario deconstructs the mainstream view of the Neural Model as an average idealized speaker by proposing a way to simulate and analyze a population of artificial individuals. We show how the amount of "shared linguistic knowledge" across speakers is highly dependent on the specific linguistic background of each individual.

Overall, we believe our framework opens the path for future discussion on the role of computational modelling in usage-based linguistic theory and vice versa, and provides a new formal methodology to both fields of study.

PUBLICATIONS

Some of the material and figures have appeared previously in the following publications:

Ludovica Pannitto and Aurélie Herbelot (2020), "Recurrent babbling: evaluating the acquisition of grammar from limited input data," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, pp. 165-176, https://www.aclweb.org/anthology/2020.conll-1.13

Ludovica Pannitto and Aurelie Herbelot (2022), "Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition?" *Frontiers in Psychology*, 13, ISSN: 1664-1078, DOI: 10.3389/fpsyg.2022.741321, https://www.frontiersin.org/article/10.3389/fpsyg.2022.741321

Ludovica Pannitto and Aurélie Herbelot (to appear), "CALaMo: a Constructionist Assessment of Language Models," in *Proceedings of the 1st Workshop on Construction Grammars and NLP*, Association for Computational Linguistics, https://arxiv.org/abs/2302.03589

The code developed is available at:

Ludovica Pannitto (2019–b), *LSTM*, https://github.com/ellepannitto/LSTM/tree/master

Ludovica Pannitto (2019–a), *CALaMo Toolkit*, https://github.com/ellepannitto/Catenae

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

1 INTRODUCTION

Over the decades, linguists have given a lot of thoughts to what language actually is, and how it can best be formally described. As a uniquely human phenomenon, language is extremely multifaceted: different approaches and theories with different aims have produced an extremely rich array of conceptual tools that can be used to describe language in its different aspects.

Computational modelling has largely been used to simulate and investigate aspects of languages at various levels of granularity and is also becoming more and more crucial as a reverse-engineering approach to tackling known questions in cognitive science or psycholinguistics (Dupoux, 2018). A specific area is concerned with **Language Modelling** (see Section 2.1), namely the reproduction of linguistic surface structure by means of probabilistic models. **Neural architectures** have played a special role in this subfield of research, thanks to their enormous flexibility.

The extremely varied conceptual complexity found in linguistics, that is, in the study of language as a human means of expression, gets cut down by order of magnitudes when it comes to the analysis of language processing in computational modelling. In the realm of neural language modelling, in fact, whether systems acquire any sort of linguistic knowledge remains one of the biggest conundrums of the field.

When *language* is mentioned in relation to Artificial Neural Networks (ANNs), it mostly seems like the word is used as a perfect synonym of *grammar*: while it is clear from the theoretical perspective that the two objects do not entirely overlap, it seems that the distinction gets blurred in the computational arena. This way, a lot of assumptions that are made when abstracting away from the language level on to the grammar level, usually clearly stated or implied in linguistic research, are not recognised as such in computational fields. We will in fact argue that it is often the case that a **specific set of choices** concerning the description of language are taken as **default**.

Most current work seems to implicitly make a number of assumptions about what kind of grammar is supposed to emerge from neural language models, and this underlying choice is often echoed in the

most common evaluation settings and in the conclusions that are being drawn from such experiments. As we will make more explicit in Section 1.1, most of these default assumptions are inherited from the **nativist Chomskian tradition**, which has pervaded a lot of the computational work on grammar, and continues to do so in the recent literature on neural models. In our literature review (Chapter 2), we will talk in more details about the specific postulates that have been integrated into current frameworks, and whether this integration was warranted, given the architecture and learning behaviour of neural models.

From a linguistic perspective, the fact that, when it comes to computational models, so many nativist assumptions permeated into the mainstream methodology seems at odds with the very nature of those computational models. Neural models are in fact essentially based on **pattern learning** and completely agnostic about the nature of the data they are made to process. The idea that language can be abstracted from a general purpose statistical mechanism is more akin to **usage-based approaches**, and neural language models would provide a much more natural testbed for theories of that kind.

In one of the foundational works of the usage-based theorization, for instance, Tomasello (1999) highlights how the development of language skills can be considered a broader process of adapting to acquire cultural knowledge in various domains, and that pre-existing mechanisms such as schematization, categorization, statistical learning and analogy-making, already present in primates, determined the grammaticalization of linguistic structures and are enough to do so on the ontogenetic level, too.

In the cognitive and usage-based accounts, moreover, the exploitation of **predictability** during language development (and again we refer to development at all the three tiers of philogeny, ontogeny and cultural evolution) is the root of a number of fundamental mechanisms such as schematization, entrenchment and distributional analysis (Lewkowicz et al., 2018). By these processes, language, namely a structured inventory of constructions, gets built through generations and throughout a speaker's lifetime. In this framework, for instance, shared linguistic material among utterances, such as morphological markers, enable the identification of particular patterns or constructions as units bearing meaning (Croft, 2001).

This perceived gap between the nativist and non-nativist tradition with respect to computational modelling probably mostly stems from the fact that different theories have emerged from different communities, and only some of them have co-evolved with the computational modelling community: in the wide array of approaches to linguistic studies, the Chomskian school offered a definition of language that could most easily be computationally interpreted and implemented, thanks to its formal approach. Hence, it has gained a more central role

within the computational community than different approaches.

The purpose of this work is to look at the current mainstream training and evaluation methodology for **Neural Language Models**, taking into explicit consideration the linguistic assumptions made along the way. Throughout this thesis, we will take the idea of a 'model' seriously, i.e. we regard the computational framework as a way to simulate some part of reality in a vacuum, making explicit simplifying assumptions in the process. We argue for a tighter **integration** between the description of such a model and the linguistic theory that it is supposed to simulate. That is, we would want to see how the working parts of a model encode specific theoretical statements, and most importantly, we would like to know where the model simplifies and where it directly contradicts the theory. To that end, we will propose an improved methodology to tie up computational modelling and linguistic frameworks, and we will test this methodology in two actual case studies relating to the grammatical abilities of Recurrent Neural Networks (RNNs).

1.1 NATIVIST VS. NON-NATIVIST APPROACHES TO LANGUAGE ACQUISITION

People use language creatively. This ability to manipulate conceptual units, despite seeming a very superficial, maybe even naive and intuitive aspect of the human linguistic ability, is actually at the core of many properties that natural language exhibits and should be taken as both the starting point and the guiding light of any theory aimed at explaining how natural language, broadly speaking, develops.

Creativity, which we simply define as the ability to reuse existing, small linguistic components to build up new, unseen blocks, has been in fact mentioned as one of the traits that best distinguish human language from animal communication systems, and, more strikingly, it has also been recognized as a skill that speakers acquire over time (Bannard et al., 2009): **linguistic productivity** is gradually acquired by children, with competence building up on knowledge about specific items and on restricted abstractions before, if ever, getting to general categories and rules (Goldberg, 2006; Tomasello, 2003).

All theories of language development and use recognize that at the root of human linguistic ability is the capacity to handle symbolic structures: what theories do not agree on is the **content of people's linguistic knowledge**, how this content is acquired and to what extent linguistic creativity is affected by this stored knowledge (Bannard et al., 2009).

In recent formulations of the Universal Grammar (UG) framework (Hauser et al., 2002), the child's linguistic knowledge is described

in terms of abstract rules and categories. But many studies have questioned this assumption, showing how the empirical input to which children are exposed is enough to explain much of their linguistic development, provided that the child is equipped with the right tools to decode it.

Such studies, that broadly fall under the category of **usage-based models**, have argued against the two main tenets of generative models, namely the *poverty of the stimulus* (Chomksy, 1959; Chomsky, 1968) and the *continuity assumption* (Pinker, 1996): usage-based theories focus instead on the ways in which language can be represented as a rich-enough signal for learners to pick up on, as well as on the cognitive mechanisms of **attention** and **memory** that explain and constrain many phenomena in language learning.

The issue of tracking statistics in the input has been reasonably settled (R. L. Gómez and LouAnn Gerken, 2000; Saffran, Werker, et al., 2006), and interest has shifted to a whole array of new issues concerning how children use the acquired **patterns** (Romberg and Saffran, 2010) and about the nature (Perruchet and Pacteau, 1990; Perruchet, Vinter, et al., 2002) and content (Estes et al., 2007; Chen Yu and Ballard, 2007) of the representations, as well as the nature of the learning process itself. We examine these topics further in the following section, starting with the broad question of the **input** (Section 1.1.1), and dedicating the remainder of the section to two equally foundational aspects: the approach to **stability** (Section 1.1.2) and the already mentioned ***continuity hypothesis***, and **systematicity** as the defining aspect of grammar (Section 1.1.3).

1.1.1 *The input*

One of the main arguments introduced in the nativist framework is that of the *poverty of the stimulus*: the input to which children are exposed is underdetermined and does not explain acquisitional generalizations that learners are able to perform (Crain and Pietroski, 2001).

The idea that the input is only marginally relevant to language acquisition has had solid effects in the generativist tradition, which has generally paid less attention to the acquisition process and mechanisms themselves and to the analysis of the peculiar features of child-directed language. It has also resulted in little attention being paid to the mechanisms that could enable children to *extract* linguistic structure from the linear signal.

Generally, generativist theories assume that children navigate an **hypothesis space** that is defined by innate constraints (Eisenbeiß, 2009): on the basis of evidence, various algorithms have been proposed for children to be able to select the right hypothesis. Learning is assumed to involve two different and rather independent aspects:

acquiring the lexicon and all the peripheral aspects of language, and setting innate parameters to the right target values. Note that this process does not require any hypothesis formulation step, as the options are readily available in the learner's brain.

In contrast, constructionist perspectives propose that language arises from **domain-general cognitive processes** that interact with input, suggesting that the input is tailored and biased to facilitate the learning process (Boyd and Goldberg, 2009).

Extensive research has demonstrated that children have strong abilities in statistical learning. More complete reviews have been published by R. L. Gómez and LouAnn Gerken (2000) (first studies in statistical learning), (Romberg and Saffran, 2010) (with a specific focus on first language acquisition) and Christiansen (2019) (different approaches to *implicit statistical learning*).

Infants and young children have been showed to learn language through statistical learning mechanisms, relying on statistical relationships between speech sounds to isolate word chunks (Saffran, Aslin, et al., 1996) and acquire grammatical information (Gomez and LouAnn Gerken, 1999). These abilities are not limited to linguistic contexts and can also be observed in non-linguistic contexts (Lewkowicz et al., 2018). Evidence for language acquisition through statistical learning is instead more mixed when it comes to detecting non-adjacent structures and more complex patterns.

The crucial difference between the nativist and the non-nativist approach here is how strict the relation between the received input and the acquired linguistic structure is: if we commit to a view in which the input only serves as a **trigger** of an almost pre-determined cognitive structure, we are naturally driving our attention far from the features of the input and primarily to the features of the structure. On the other hand, deriving the linguistic structure from the **input structure** itself imposes the necessity to look at the two aspects together.

1.1.2 *Stability: continuity hypothesis and the native speaker construct*

The *continuity assumption* was first introduced in Pinker (1996) in order to approach developmental language in a framework that was actually developed to analyse adult language (Tomasello, 2003), thus relying on two main assumptions:

- differences between the adults' and children' linguistic structures are negligible;
- observable or surface differences are just due to performance factors.

Different models, based on the so called *developmental hypothesis*, have emerged to counter the generative approach: the mechanisms

underlying acquisition remain the same throughout the life-long acquisition process, but the structures and abstractions produced by them evolve throughout the different stages. This leads to the other aspect that distinguishes the usage-based approaches from the generative ones, namely the emphasis that usage-based models put on the **time-dependent** nature of the linguistic signal. While certainly not denying the utter relevance of hierarchical structures in language comprehension and production, they advocate that it emerges from the fact that language must be processed linearly and is subject to constraints imposed by general-purpose memory and cognitive mechanisms (Christiansen and Chater, 2016b; Cornish et al., 2017). The existence and facilitatory role of higher-order structures is unquestioned and consistent with general observations about memory, such as the well-known constraints on our ability to recall stimuli (Miller, 1956).

Both connectionist (Elman, 2001) and constructionist approaches (Goldberg, 1995; Tomasello, 2003) can be said to follow the *developmental hypothesis*.

According to the usage-based account, for instance, generalizations come into being gradually, as the progress to **productivity is gradual** stemming from item-specific knowledge (Bannard et al., 2009): the evidence about a child's own knowledge of grammatical structure is in fact contradictory (Dittmar et al., 2008; Gertner et al., 2006).

Dabrowska (2015) provides counterarguments to most of the assumptions made by universal grammarians. Specifically with respect to the *continuity hypothesis*, she points to a number of studies that provide evidence for discussion against some widespread Universal Grammar arguments. As she points out, mere exposure is not enough for language acquisition and the child needs to find themself in an interactive environment (Sachs et al., 1981; Todd and Aitchison, 1980). The course of language development seems to be quite different from child to child, also depending on how individual learners *break into* grammar.

Most importantly, Universal Grammar posits that all speakers eventually converge to the same grammar (Crain, Thornton, et al., 2009; Lidz and Williams, 2009). **Individual differences** however have been found in almost every area of grammar, depending on a variety of factors including environmental ones (Street and Dąbrowska, 2010).

It is again a matter of framing: as Dabrowska (2015) points out, individual differences are not in principle incompatible with innateness, as they could be just as well due to inheritance factors. We do not, however, necessarily want to assume continuity and stability when it comes to the description of grammar itself.

Individual variation is similarly suppressed by the theorization, again typical of universal grammar approaches, of the **idealized native speaker**: the almost exclusive focus on linguistic competence has automatically implied the exclusion of non-native proficiency from the scope of investigation (Radwańska-Williams, 2008). This is again due to the fact that in the Chomskian tradition the speakers' community is viewed as **homogeneous**, as everyone converges to the very same grammar: linguistic knowledge is equally possessed by any component of the community. As Radwańska-Williams (2008) points out, the idea of a linguistic community as an homogeneous system stems from the structuralist tradition, where language was described syncronically as the result of the aggregation of different realizations in the speech community. With Chomskian tradition, this homogeneity is however **projected** on the single speaker: what was originally conceptualized as a social dimension has thus become an individual feature. Leaving aside the political implications of the concept, that Radwańska-Williams (2008) however introduces, the linguistic point remains: there exist huge variations in competence within a speech community, and non-native speakers can reach levels of proficiency that can be considered similar to those of the rest of the community, and their linguistic productions however contribute to the definition of the **linguistic community**.

1.1.3 *Systematicity: compositionality vs productivity*

Systematicity is arguably the most desirable skill that a Language Model should acquire, in order to say that it has acquired language as a whole. The ability to understand and generate, with finite means, an unbounded number of novel sentences is in fact universally considered one of the hallmarks of our language faculty. This unboundedness is however not completely arbitrary, despite being systematic. The boundaries of this *systematicity* remain largely unclear: provided that we agree on what the finite means at our disposal are (and we will argue that this basic difference between theories is what determines the difference in focus), not all the possibilities are actually realised by speakers and not all the realised possibilities share the same cognitive or linguistic status.

One way to look at *systematicity* is that of **compositionality** (Goldberg, 2015; Aurelie Herbelot, 2020; Partee, 2004), usually referred to in one of the two classic forms, both often attributed to Frege:

BOTTOM-UP COMPOSITIONALITY *the meaning of the whole sentence is a function of the meanings of its parts*

TOP-DOWN COMPOSITIONALITY *words have meaning only as constituents of (hence, presumably, only in virtue of their use in) sentences*

Both versions of the principles are strongly radicated in the Chomskian nativist theoretical asset. The most widely known version of compositionality is probably due to Katz and Fodor (1963), that port Chomsky's innateness theory to semantics: a set of rules or constraints is needed in order to systematically build the meaning of sentences by integrating meaning of words.

Even the Montagovian formal approach to compositionality (Montague, 1970; R. Montague, 1970) relies on Chomskian-derived ideas of a **stable lexicon** that stores meanings and the existence of a set of precise **interpretation rules** that allow for those meanings to be mixed and modulated *through* the filter of syntax.

Despite dealing with semantics, the core of both visions is still very much syntax-centered, to which semantics has to be isomorphic in order to function. Moreover, very little space is left for indeterminacy, negotiation between speakers and other aspects related to the interactive and communicative nature of language (different individuals can retain in fact quite different concepts associated to the same lexical label for instance, Labov, 1973).

In other words, if we see systematicity from the standpoint of compositionality, the quasi-regularities of linguistic structure represent a major hurdle to surpass.

Quasi-regularity or quasi-systematicity is instead the rumbling engine of **productivity**, as in the ability of speakers to use all the available linguistic means to cue the intended meaning.

Just like compositionality, productivity deals with the domain in, or the extent to, which a grammatical pattern can be employed in a linguistic context without losing interpretability. And, just like compositionality, it deals with what is actually possibile in the language and where so draw the boundaries of acceptability or interpretability.

The shift has not been just syntactic: in the formal representation of these two aspects of systematicity in semantics, for instance, composition-oriented or productivity-oriented theories have conceptualized the idea of selectional constraints differently. In Katz and Fodor (1963) for instance the idea is that formal constraints are hard-coded in the lexicon, in order to regulate ambiguity and semantic acceptability. Later on, for instance in Fillmore (1976)'s model, selectional constraints are relaxed to selectional preferences, making space for **individual manipulation** in rules and boundaries.

Knowledge on *systematicity* is in both cases considered as implicit knowledge that the speaker has about their language. Differences in the theoretical accounts naturally entail that the focus of the two approaches is not the same. Nativist approaches have in fact primarily dealt with compositionality: given grammar rules and lexicon, how do these combine to convey the intended meanings? Usage-based theories, on the other hand, have primarily been dealing with productivity: how

far can meaning boundaries be forced? What are the mechanisms that allow for linguistic creativity? This of course entails, in the usage-based community, a relation to surface properties of the input as well: Croft and Cruse (2004), for instance, note how the maximally schematic constructions, such as `sbj verb obj` for instance, are also the most productive ones, and that this has a relation to their frequency too, both as a type and for each of their instantiations.

1.2 NEURAL LANGUAGE MODELS IN LINGUISTICS

Over the last decade, Neural Networks have become the *de facto* standard for computational approaches to language modelling. Different architectures have been appearing and outperforming what was previously considered state-of-the-art on what looks like a fast moving trademill.

When I began the work on this thesis, in 2018, the turmoil around the now state-of-the-art Transformer Models had not yet fully exploded and Long Short-term Memory Networks still seemed the most reliable option for Language Modelling. Today, four full years later, we have discussed and learned a lot more about transformers architectures: for the reasons that will be made clear in the remaining of this paragraph, we however chose not to consider them in our experiments, as they do not fit the desiderata we are looking for in a linguistic simulator.

Before discussing the experimental setting, in fact, we should clarify the role that Neural Language Models have in our approach. More specifically, we do not ever consider them as cognitive models: in other words, we are not arguing for any kind of architectural similarity between the human cognitive processes and the computational tools.

What we are supporting is a kind of **behavioral similarity**: the model builds some sort of abstraction over linguistic data, and based on this abstracted representation it responds to our stimuli. The abstraction built by the model is linguistic in nature, and can therefore be used to explore the boundaries of the language system similarly to what can be done with human subjects.

1.2.1 *The effect of choices on Neural Language Model studies*

The perspective taken on each of the aspects mentioned in Section 1.1 has, we believe, cascading effects on the experimental asset and the conclusions drawn from the Neural Language Model's responses to our setting.

INPUT As we mentioned already, the difference relies on whether we consider the input as determinant to shape the learner's grammatical knowledge or not. In the first case, attention will be drawn to the processes that allow to learn from positive evi-

dence, and the strict relation between the abstract grammatical structure of the input and that that will be acquired and therefore produced by the learner. In the latter case, the input will only serve as a triggering factor and its features will therefore play no role in the analysis;

STABILITY Depending on the view that we take on the *continuity hypothesis*, we could see Neural Language Model's grammatical competence either as a switch or as a gradient property. In the first case, we will test whether the network is able or unable to handle some linguistic phenomenon, while in the second case we would be interested in seeing how and why some linguistic aspect becomes more and more salient to the network during training, and which factors determine the acquisition of a certain pattern;

SYSTEMATICITY Committing to the compositionality or productivity perspective entails a different organization of linguistic knowledge: the compositionality perspective tends to set meaning aside, and treat the lexicon as an organized repository of semantic information, while grammatical knowledge is made up of rules. In this setting, it makes sense for instance to test Neural Language Model's capabilities on semantically nonsensical sentences or to extend the known rules to completely unknown lexical items. In the productivity perspective, instead, meaning is intrinsically part of the process and is treated as a systematic aspect of grammar, too.

As our aim is to explore and control for these mentioned aspects (i.e., *input*, *stability* and *systematicity*), we do not include Transformer models in our work. Transformers are in fact the most widely employed architecture in Large Language Modelling and they are currently showing impressive results both in terms of surface patterns and performances on downstream tasks. They however need larger amounts of data and a pretraining phase, which makes it virtually impossible to control for input conditions. Similarly, the existence of such pretraining step defeats the purpose of investigating *stability* as knowledge builds up on pre-acquired biases. However, as we will discuss in Chapter 2, more and more interesting works are emerging in the field and they constitute the current standard in language generation tasks.

1.3 OUTLINE

Having highlighted the important methodological issue related to the use of computational modelling for linguistic research, we now summarise how we will tackle the problem throughout this thesis. In the first part, we will deal with theoretical aspects, reviewing the

recent literature in the light of various linguistic assumptions and showing how a model statement would clarify such assumptions. In the second part, we will propose a case study consisting of two specific experiments, and show how explicit model statement helps analysing and interpreting the results of those experiments.

We expand the proposed structure below, in the form of three main research questions.

1.3.1 *Q1: what are the assumptions made by current literature and in which way can they be theoretically harmful?*

The first part of the thesis (Chapter 2) introduces our methodological claim, by first providing a focused review of the recent literature on computational modelling of language acquisition, specifically focusing on LSTMs as basic associative models and their relation to general-purpose human learning processes. We start with the observation that most work in this area presupposes a number of features in the emergent grammar, which actually correspond to a certain take on language development (namely, the Chomskian tradition). Breaking down the existing work in terms of those assumptions, we also suggest that an alternative reading could be developed, based on a competing theoretical approach: usage-based theories. We specifically highlight how the evaluation of computational models becomes biased due to a lack of explicitness in relating experimental and theoretical aspects of the research question. Such bias is evident at the quantitative level (i.e. the chosen evaluation measures may push interpretation in a certain direction), but also in the analysis of results, where discussions are often marred with implicit postulates.

1.3.2 *Q2: Can we make explicit the relation between theory and model abstraction?*

In the following chapters (Chapter 3 and Chapter 4), we provide a first case study which explores the relation between particular linguistic issues and aspects of the model. We first develop the idea of '*model abstraction*', i.e. the conceptual structure and dynamics associated with a given algorithm. This includes things like the nature of the representations in both input and output, as well as the specific learning mechanism at work.

We then specifically investigate two questions:

- the nature and role of the input data in computational language models, and its relation to theoretical insights on child-directed speech or innateness;
- the nature of the latent linguistic representations that emerge in the course of learning.

The role of input

The first question stems from long-standing questions in the linguistic literature. In particular, how does language influence abstraction and allow you to get to adult-like grammar as quickly as possible? Artificial Neural Networks are typically trained on cognitive implausible input, that does not resemble the kind of language children are exposed to during development: among the shortcomings of the studies in literature one must also consider the fact that the majority of the claims have been made on models that have only seen a very specific and constrained variety of language, that does not resemble the one learners grow up with (E. V. Clark, 2009; Snow, 1972). To alleviate such issues, we propose that, by varying the quantity and quality of the input data, a more precise picture could be obtained with respect to what the patterns learned by the LSTMs look like, with reference to the different assumptions that theories of syntax acquisition make. While the aim is certainly not to reproduce the linguistic environment to which the child is exposed and neither to mimic their cognitive development, as too many other and possibly extra-linguistic factors come into play during acquisition (e.g. pointing, joint attention, intention reading... Tomasello, 2003), it seems nonetheless interesting and feasible to investigate through the Artificial Neural Networks framework which features of child-directed language are the ones fostering the grammatical abstraction process.

Our results confirm that Neural Models are proficient at reproducing statistical regularities found in the input and do so beyond the lexical level.

The nature of latent representations

Our second investigation concerns the nature of the linguistic representations that can be seen emerging in the process of language development. We have already mentioned earlier that statistical models without hard-coded, 'innate' knowledge, might be better suited to investigate usage-based theories of language than Chomskian approaches to language development. We take this idea further and ask whether the grammatical knowledge acquired by Recurrent Neural Networks might be analyzable in terms of constructions.

Working in a usage-based framework implies a shift in considering the relation of lexical bias to syntactic abilities: a substantial amount of previous work seems to make strong assumptions in that regard, such as the fact that the network acquires grammatical rules which are then applicable to whichever lexical item. For instance, most approaches will test whether the network has acquired 'agreement', as if 'agreement' were a structure that fully emerged at some point during processing, independently from the features of the language fed to

the network, and did not evolve. In those accounts, grammatical structures are then regarded as *parameters* that, once set, are conceived as stable and valid. This is in contrast with theoretical frameworks that hypothesize an evolution in grammatical abilities. We further analyze to what extent the learning process leading to abstraction can be said to be incremental.

Our main contribution consists in showing that a latent relation exists between distributional meaning and the way schematic patterns are reproduced by the network in the course of learning.

1.3.3 *Q3: Language and the boundaries of variability among speakers*

In the last part of this thesis (Chapter 5), we will consider the issue of evaluation, keeping in mind the theoretical arguments made in the previous sections. We will start from the observation that a large part of the computational linguistics literature performs evaluation according to a single '*gold standard*' per task. For traditional tasks such as sentiment analysis or word similarity ratings, the annotations of human subjects are somehow averaged, and the system is evaluated against that average. For language modeling, model perplexity is computed with respect to the statistical features of a usually large corpus, which aggregates the writing styles and linguistic habits of thousands of speakers. While this state of affairs has started to be criticised by various researchers (Jolly et al., 2021; Silberer et al., 2020), it remains for now the status quo.

When considering language development as a speaker-dependent process, strongly affected by the nature of the input, an evaluation based on an '*average speaker*' becomes truly unsatisfactory. We cannot assume the existence of a ground truth, and must rely on softer evaluation measures: it is clear that the linguistic behaviours of different speakers must overlap sufficiently to allow for communication (Marti et al., 2019; Pickering and Garrod, 2006; Van Deemter, 2010), but that we also want to observe in the output of the network the kind of variability that is seen in humans. In particular, we might want to reproduce variations in competence at different stages of life, and for different types of socio-economic status, as well as uncertainties in a single speaker's judgements. We would also like to identify the locus of such variations, under the assumption that some 'core' constructions *must* be shared by all individuals, while others are less important to successful communication.

Such considerations naturally bring in a larger theoretical question: does language, as we formally describe it, exist in any actual speaker? It is in fact likely that the models found in the linguistic literature capture commonalities across individuals, rather than the reality of each subject. This is in fact explicitly stated in the famous Chomskian

distinction between *competence* and *performance*. For Chomsky, competence is the ideal knowledge of one's language, while performance is the faulty, observable usage a speaker makes of it. One could reframe this distinction by saying that competence is what is *core* to a language – and thus shared across a community – while performance denotes non-core, potentially idiosyncratic aspects of an individual's linguistic knowledge.

The last part of this thesis proposes a framework where acquisition is modelled across a range of (artificial) speakers, trained on different data, rather than in a single individual. The idea is that by explicitly generating a number of model instances, we can perform analyses that start elucidating potential variations in linguistic knowledge, and assess whether the shared constructions across speakers reflect what we might otherwise understand as 'core competence'.

We therefore tentatively provide a definition of *language* as *shared linguistic knowledge* in a community and provide a novel approach to model the speaker-listener interface.

1.4 CONCLUDING REMARKS

This work was developed in the context of a Cognitive Science department. So, what has all of this to do with Cognitive Science? A core issue concerning language development is the nature of **linguistic representations**, and consequently the choice of a linguistic theory to describe and formalize them. We feel this aspect has been overlooked in the Neural Language Models literature and we attempt an approach that brings back linguistic theory into the picture, as a bridge between Natural Language Processing and Cognitive Science. To this end, we are committing to the usage-based constructionist theoretical framework: we are not debating whether this represents a better model for human language acquisition, but rather whether the tools and categories introduced by construction grammar can be enough to explain Neural Language Model's produced language.

From the Cognitive Science perspective, one of the major issues in the study of language acquisition is the nature of speakers' underlying linguistic representations, and their **development** during their lifetime. Since learning a language largely overlaps with **learning how to process the input**, there must be a relation between processing biases relating to certain types of constructions and the distribution of those constructions in the linguistic input (Christiansen and Chater, 2016a). As **experience** grounds linguistic knowledge, **distributional properties** constitute a key aspect to determine the content of linguistic representations. Language is not in fact considered as an autonomous cognitive system, but rather the acquisition of grammar is regarded as

any conceptualization process and knowledge of language as **knowledge** in general, therefore emerging from use (Croft and Cruse, 2004).

Nativist theories, which typically do not posit a tight relation between language processing and the acquisition of linguistic categories, have overwhelmingly represented a fundamental approach to language sciences, thus greatly contributing to the understanding of aspects of both language and the mind. However, their legacy has perhaps spread outside of the boundaries and assumptions that were posited by Chomsky and his entourage. As Christiansen and Chater (2016a) note, isolating the study of language from considerations regarding processing, acquisition and evolution has affected the way researchers have approached the observation of linguistic phenomena outside of the Universal Grammar theory *stricto sensu*.

In the usage-based tradition what determines representations is surface properties such as frequency of occurrence and meaning: distributional properties of the utterance can therefore be taken as proxies to cognitive representations.

2

RECURRENT NEURAL NETWORKS AND USAGE-BASED THEORIES OF GRAMMAR ACQUISITION

The content in this chapter partially appeared in: Ludovica Pannitto and Aurelie Herbelot (2022), "Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition?" Frontiers in Psychology, *13,* ISSN: *1664-1078,* DOI: *10.3389/fpsyg.2022.741321, https://www.frontiersin.org/article/10.3389/fpsyg.2022.741321.*

Artificial Neural Networks, and Long Short-Term Memory Networks (LSTMs, Hochreiter and Schmidhuber, 1997) more specifically, have consistently demonstrated great capabilities in the area of Language Modelling (Section 2.1). In addition to generating credible surface patterns, they show excellent performance when tested on very specific grammatical abilities (Gulordava et al., 2018; Linzen and Baroni, 2020), without requiring any prior bias towards the syntactic structure of natural languages. As Linzen and Baroni (2020) point out, such successes invite a **reassessment** of classic arguments that language acquisition necessitates rich innate structure (Chomsky, 1968; Pinker, 2009).

However, an account that links Neural Language Modelling to usage-based linguistic formalisms is still missing. This is not entirely surprising, as Neural Language Models have developed mostly independently from research on infant language acquisition. But it leaves fundamental questions open, such as how and to what extent Neural Language Models can be used to simulate the basic, domain-general mechanisms driving the acquisition process, as posited by usage-based theories. More generally, at the present time, many results in the Neural Language Modelling literature stay controversial, and the picture is not yet complete: it remains unclear how and to what extent grammatical abilities emerge in artificial language models, and how this knowledge is encoded in their representations.

It is our belief that part of the controversy stems from the **discrepancy** between the theoretical and the computational field, when it comes to the discussion of fundamental linguistic concepts. Current computational approaches to the evaluation of Neural Language Models' grammatical abilities take implicit stances on what *human-like* generalization should look like: for instance, not enough importance is given to the input on which Neural Language Models are trained, as this is probably seen as an epiphenomenon of a much more general language competence. Moreover, most studies consider grammatical abilities as a separate phenomenon from lexical biases, often evaluat-

ing the network on discrete generalization (i.e, the network has either acquired the rule or it has not). Their results are therefore hard to compare to psycholinguistic results on the child language acquisition process, especially when experimental evidence comes from usage based approaches (Langacker, 1988) which grant the input a much bigger role (Boyd and Goldberg, 2009). Similarly, other aspects end up often being overlooked, as they can be hard to frame in the mainstream approach, which is largely influenced by nativist assumptions. We find it essential to remark the fact that we do not find nativist assumptions erroneous in any way, nor do we intend to show the primacy of one specific theory over the others. Our only aim is to highlight the fact that there exist a number of latent biases in the approach that the computational community currently has when discussing Neural Language Models' linguistic performances: acknowledging the existence of such biases can lead to the definition of better frameworks and evaluation benchmarks, and ultimately to a cleaner theoretical discussion.

Sorting out the relation between computational modelling and theory is especially essential as computational methods are becoming more widespread in the scientific study of language. As pointed out by Dupoux (2018), "*reverse engineering* language development can contribute to our scientific understanding of early language development": computational modelling can offer promising features of scalability and reproducibility, in an environment where different input-specific features are easily isolated and evaluated. The characteristics offered by the computational environment come in handy when we deal with aspects that are notoriously difficult to isolate or estimate: evaluating the influence of input-specific features on language learning is one of these cases, where setting up a real-life scenario is especially hard.

But in spite of the advantages of computational experimentation, their output can only be taken seriously at the point where hypotheses are situated in a relevant theoretical framework. This chapter attempts to clarify the relation between modelling and theory at the current point in time. It starts with a review of Language Modelling as a computational practice (Section 2.1), briefly describing common evaluation methodologies. It then goes into further detail in Section 2.2-Section 2.4, reviewing how the literature so far has treated the specific aspects of *input*, *stability* and *systematicity*.

2.1 LANGUAGE MODELLING

Language Modelling can be defined as the task of determining the **probability** of a given linguistic sequence, through statistical analyses of a body of data. Modulo the due differences, language modelling touches one of the most fundamental debates concerning the acquisition of linguistic abilities by human speakers. In order for a compu-

tational model to be able to perform a task of this kind, in fact, it is assumed that many aspects of language ought to be mastered: all linguistic levels are involved when predicting a word in context (i.e., the prediction has to be morpho-syntactically accurate, semantically acceptable, but also, for instance, pragmatically appropriate). Therefore, mastering the language modelling task can, at a very coarse-grained level, be considered as difficult as mastering language itself.

In their simplest flavour, language models come in the form of sets of *n-grams*, i.e. sequences of n words, where n can be regarded as the amount of **context** to be considered. The model then assigns probabilities to longer sequences based on the estimated probabilities of n-grams given a certain corpus. Despite their simplicity, *n-gram models*, with the appropriate context size, already show quite impressive modelling capabilities.

The advent of neural approaches in computational linguistics further changed the approach to Language Modelling as well, going from a strict statistical/probabilistic perspective to a more continuous and distributed view of language representation. Modern language models are now built on a variety of tasks that go well beyond the simple item-in-sequence prediction (some will be detailed in Section 2.1.1): while their full potential is yet to be discovered, and despite the many criticisms that have emerged in the literature (Section 2.1.2), this also allowed for different linguistic aspects to be encoded in the representations, thus making it possible to investigate linguistic aspects beyond the sequential, morpho-syntactic level.

Recent experiments, including one conducted by Cornish et al. (2017), have demonstrated the emergence of language-like structure from linear signals: the authors demonstrated how cognitive limitations of human learners may lead to adaptations that result in important aspects of the sequential structure of language, such as its characteristic reusable parts. In a letter-string recall task, participants were asked to reproduce a series of 15 string that they had been previously been trained on. The recalled strings were used as inputs for the next participants, in a series of 10 subjects for chain. The authors report that, across generations, not only does **learnability** increase (i.e., the overall accuracy of the recalled items in terms of normalized edit distance increases, and not at the cost of a collapse of the string sets into very short sequences), but the amount of **reuse** of chunks also significantly differs from what one would expect from random strings, and structure similar to natural language generally emerges. In order to determine the increase of distributional structure, Cornish et al. (2017) adopt a metric which is frequently used in artificial grammar learning studies: *Associative Chunk Strength* (ACS) (Knowlton and Squire, 1994): for a given test sequence consisting of x bigrams, and $x - 1$ trigrams, ACS is calculated as the relative frequency with which those chunks occur in the training items. For example, ACS

for the recalled item ZVX in generation t is calculated as the sum of the frequencies of the fragments ZV, VX and ZVX in generation $t-1$ divided by 3. By means of averaging, the authors find that the next generation tends to reuse these chunks successfully, and more so as generations proceed, thus incrementally developing re-usable units.

At a general level, the parallel between human processing of linear signals and language modelling is clear. But at a deeper level, it is worth remembering that, despite their impressive results, language models and Neural Language Models in particular cannot be considered cognitive models. The history of psychologically plausible computational modelling goes in fact well beyond artificial neural networks: some examples can be found in Freudenthal et al. (2015), McCauley and Christiansen (2019), and Solan et al. (2005) for instance. Still, while a structural parallelism is impossible to claim (and not what we are aiming for in this thesis), neural models embody some of the **functional principles** at the core of usage-based theories, and their mechanisms can be powerful tools to explore the effects of specific features of the input language, in a way that is hard to replicate in real life scenarios. For these reasons, they will be the default computational model for our work.

2.1.1 *Architectures*

Two specific architectures have proved to be particularly suited for the language modelling tasks: recurrent architectures such as LSTMs (Hochreiter and Schmidhuber, 1997) and more recent models based on so-called Transformer architectures (Brown et al., 2020; Devlin et al., 2019; Radford et al., 2019a).

LSTMs are simple models, that can reach reasonable language performances, and whose features mirror at least in part the general-purpose mechanisms advocated by usage-based models of human language acquisition. Their introduction is regarded as a milestone for language modelling, as they reached unprecedented generalization capabilities, and on the other hand they are rather simple architectures with respect to more recent models. They can still be trained with few resources (in terms of space, amount of training data needed and computational resources) and therefore are suitable tools for research in language acquisition.

LSTMs have been applied, without substantial modifications, to a variety of tasks, ranging from time series prediction to object co-segmentation, and encompassing grammar learning as well. On the continuum between specialized devices and **general purpose associative mechanisms**, LSTMs place themselves on the latter side, with their recurrent structure seeming to be crucial in the linguistic abstraction process (Tran et al., 2018).

LSTMs are usually trained on Language Modelling tasks: predict the next word given the previous sequence. While in a simple feed-forward neural network, the prediction at time t depends on the input at time t alone, in a recurrent neural network such as an LSTM, contextual information is maintained from one prediction step to the next. The output of the network at time t depends therefore on a subset of the inputs fed to the network in a set window. With respect to standard Recurrent Neural Networks, while being trained on a language modelling task, the LSTM also tunes itself to a specific time dependency distance, learning what to remember and what to forget. From a linguistic perspective, the sequence-dependence of LSTMs prediction make them great tools to simulate language processing that obviously involves the understanding of a time evolving signal.

Transformer Language Models (TLMs) are, on the other hand, non-recurrent architectures and are not trained on language modelling tasks in the same way as LSTMs, but rather on more complex tasks that, while surely requiring linguistic knowledge to be solved, are not as straightforward as simple language modelling. BERT (Devlin et al., 2019; Elazar et al., 2020) is for example trained on the *Masked Language modelling* task (a portion of tokens in each sequence is replaced by a placeholder and the task is to predict the original value of masked tokens) and the *Next Sentence Prediction* task (predicting if, given two sentences, one follows the other or not). For GPT-2, the *Permutation Language modelling* task has been introduced, that exploits random permutations of the sentence as context to learn items' representations.

The internal structure of the two architectures is also very different, with transformers being much more complex than either simple or layered versions of LSTMs: GPT-2 counts as many as 1.5 billion parameters, while for LSTMs this number depends on design choices (for instance, for the model presented in Gulordava et al. (2018) that we will examine in the next paragraphs, the number of parameters is around 3.5 millions).

TLMs seem to have again revolutionised Language Modelling, showing extremely realistic performances in generation tasks, however at the expense of hugely unrealistic training phases and quite complex architectures and tuning procedures. The amount of input language they require (8 million pages for a total of 40GB of text for GPT-2) is unrealistic both with respect to what a human is exposed to during their lifetime, and with respect to the collectable data with respect to children language acquisition. Due to the high cost of their training in terms of computational resources (Strubell et al., 2019), they are also not easily customizable.

On transformer models and Large Language Models in general, a number of criticisms have emerged: see for instance the argumentation by Bender et al. (2021), Bommasani et al. (2021), and Weidinger et al. (2022).

2.1.2 *Analysing language models*

The analysis of syntactic abilities retained by Neural Language Models dates back quite a few years (Lewis and J. L. Elman, 2001), and many contributions on these topics have been produced lately, in accordance to the general tendency to analyze the inner-workings and knowledge acquired by neural networks (Alishahi, Belinkov, et al., 2020; Linzen, Chrupała, and Alishahi, 2018; Linzen, Chrupała, Belinkov, et al., 2019): Alishahi, Chrupała, et al., 2019, in analyzing the discussion emerged from the first BlackboxNLP workshop (Alishahi, Chrupała, et al., 2019), highlight four dominant approaches in the evaluation of NNs performances, namely:

i. **manipulation** of the input and evaluation through specialized datasets;

ii. analyses of **representations** through diagnostic classifiers or downstream tasks;

iii. **modifications** to the NN architecture;

iv. examining the performance of the network on **simplified languages** (i.e., formal languages).

These main trends in evaluating Neural Language Models' syntactic abilities entail different assumptions on the relationship between *language modelling* and *language acquisition*: from the acquisitional perspective, the first two approaches, i.e., manipulating the input and analyzing the knowledge incorporated into representations, are more easily relatable to the psycholinguistic literature.

Most of the results mentioned in this chapter focus on LSTM-based language modelling, but a very similar scientific discussion has bloomed around Transformer-based language models (Bacon and Regier, 2019; Goldberg, 2019; Jawahar et al., 2019; Lin et al., 2019). Transformer models have indeed shown to learn structural biases from raw input data (Warstadt and Bowman, 2020) and some psycholinguistic informed approaches have emerged around the architecture. But the overall field, just as in the case of LSTMs, has led to contrasting results. Related to the question of acquisition, for instance, Warstadt, Parrish, et al. (2020) and Hu et al. (2020) have compared a range of models, including LSTMs and Transformers, on different sizes of corpora. While the amount of training input clearly benefits system performance, Hu et al. (2020) also conclude that the specific hard-coded architecture of a model is more important than data size in yielding correct syntactic knowledge. Their training data is however not characteristic of child-directed input. We will discuss in the course of this chapter how this aspect may be a crucial ingredient to get consistent results in the course of analysis.

While there are signs that *some* linguistic abilities may be acquired by general-purposes language models, a proper evaluation has to consider more fine-grained theoretical details of the acquisition process. In the next three sections, we will specifically describe the state of the art with respect to the three pillars of usage-based approaches: *input*, *stability* and *systematicity*.

The next sections (Section 2.2, Section 2.3, Section 2.4) are each dedicated to one of these aspects. For each of them, we first highlight how the same aspect can be approached and interpreted differently depending on the specific point of view taken on it. We then compare psycholinguistic and computational results. In the case of *input* (Section 2.2), we also briefly describe a particularly relevant framework (namely, Distributional Semantics, Section 2.2.3) as it is directly relevant to the structure of our proposed model, described in Chapter 3.

2.2 INPUT

As far as *input* is concerned, theories place themselves on a continuum whose ends can be identified in the following positions:

- on the one end, we find the idea that the input only serves as a **trigger** of an almost pre-determined cognitive structure. This approach naturally draws attention to the feature of the structure (for instance, the nature and number of *parameters* in the Principles and Parameters theory) rather than the shape and features of the input;

- on the other end, we find theories claiming that linguistic structure is **derived** from input structure. This is the case of theories that advocate for exemplar learning, for instance. In this case, the input naturally needs to be explored and investigated alongside with the acquired and structured linguistic knowledge.

The main tenet of usage-based models states that there is a tight relation between the input and the learned representations (Boyd and Goldberg, 2009): this principle is well expressed in Construction Grammars (Hilpert, 2014; Hoffmann et al., 2013; Masini, 2016), but does not necessarily rule out the possibility of having, as part of the construction, purely grammatical items (similar to phrase structure rules) along with more meaning-filled structures (Jackendoff, 2002). In fact, some usage-based approaches explicitly highlight the idea that learning language equates learning *how to process* language, for instance Christiansen and Chater (2016c), O'grady (2005), and Tomasello (2003).

In the following, we look at the specific issue of input from three different perspectives.

1. We first review relevant psycholinguistic literature, showing that children's language acquisition process is strongly influenced by the shape of the input.

2. We then highlight ways in which language models have tackled the issue.

3. We also introduce an additional computational method known as Distributional Semantics, which is closely related to the usage-based framework in spirit, while also being suitable to the analysis of the internal states of a language model.

2.2.1 *Infants deal with complex hypotheses during acquisition*

Extensive research has demonstrated that children possess a considerable aptitude for statistical learning. More complete reviews have been published by R. L. Gómez and LouAnn Gerken (2000) (first studies in statistical learning), Romberg and Saffran (2010) (with a specific focus on first language acquisition) and Christiansen (2019) (different approaches to *implicit statistical learning*).

In linguistic contexts, already Saffran, Aslin, et al. (1996) show how infants as young as 8 months old can segment words in speech based solely on statistical relationships between adjacent speech sounds, and with very little exposure. Gomez and LouAnn Gerken (1999) show how the same limited exposure is enough, for 1-year old children, not only to acquire specific grammatical information, thus discriminating new grammatical strings from those that showed string-internal violations, but also to do so beyond the lexical level. Infants further show the ability to act rationally when multiple generalization options are available: Louann Gerken (2006) demonstrate that infants generalize based on the formal description that is more likely to have generated the input, and Louann Gerken (2010) show that just three counter-examples to a generalization hypothesis are enough to make the infant change their prediction (Louann Gerken, 2010).

The same abilities also emerge in non-linguistic contexts. In a recent study, for example, Lewkowicz et al., 2018 show that learners above 8 months of age are sensitive to the difference between hierarchical and non-hierarchical structure in the input, thus being able to generalize **recursive and hierarchical patterns** through general-purpose learning mechanisms. Similarly, Santolin and Saffran, 2019 show how predictive dependencies facilitate learning also from non-linguistic input, therefore implicitly showing the employment of a domain-general learning mechanism. Infants, moreover, seem to be able to use the provided information in rather sophisticated ways: as Goldberg, 2006 reports, in a study conducted by Gergely et al., 2002, children showed interesting imitation skills on an everyday task (i.e., turning on the light), along with the ability to perform counterfactual reasoning, of

the same kind that would be required during the preemptive process of grammar acquisition. The emergence of language-like structure from purely linear signal has also been shown in recent experiments such as the one carried by Cornish et al., 2017, that have demonstrated how important aspects of the sequential structure of language, as its characteristic reusable parts, may derive from adaptations to the cognitive limitations of human learners and users.

A crucial aspect in the detection of structural elements and the ability to abstract them in schematic patterns concerns **relational similarity**: Hudson Kam, 2009 show that statistical learning mechanisms are used by adults to track relationships between abstract linguistic categories, in addition to individual items. Evidence comes also from Bencini and Goldberg (2000): in their study, adults were asked to sort sentences based on similarity in meaning, and adults were found to be equally likely to categorize sentences based on verb meaning, a known primary predictor of sentence meaning (Healy and Miller, 1970), and based on constructional meaning, that is to say, basing their similarity judgements on relational analogies rather than lexical meaning. Markman and Gentner, 1993 had already shown this same effect in non-linguistic judgements, finding similarity judgements to be higher when representations shared the same relations among entities.

Evidence of the fact that complex hypotheses can be induced from the input comes also for aspects that have been among the strong points of nativist theories, like *auxiliary fronting*: a study by Reali and Christiansen, 2005 focuses on the specific problem of auxiliary fronting in complex polar interrogatives, showing through corpus analysis that **indirect statistical information** enables the correct placement of auxiliary words in polar interrogative sentences, and this information is enough to distinguish between grammatically correct and incorrect generalizations.

An aspect strictly related to the sequential nature of the linguistic signal is the role of **prediction**, a mechanism that is regarded as highly relevant to language processing (Pickering and Garrod, 2013), and that could also play a part in the acquisition phase: children could use predictions during conversations in order to compare their guesses with the actual received input (Ramscar et al., 2013). The role of prediction in language acquisition has recently been examined: Fazekas et al., 2020 evaluate whether less predictable (i.e., more **surprising**) input leads to more lasting change than more predictable input, finding that when a syntactic structure is presented in a surprising context rather than a predictable one, exposure to the same structure results in an increased learning rate. This is particularly relevant to us as the **error-based paradigm** can explain domain general abilities (Stahl and Feigenson, 2015) and can be easily compared to the language modelling algorithms described above.

2.2.2 *Neural Language Modelling and input*

In Neural Language Models, the tendency so far has been to extensively evaluate systems on **acceptability** judgements. For instance, Hu et al. (2020) present a test suite including an enormous variety of syntactic phenomena, some explicitly modeled from an introductory syntactic textbook. Generally speaking, their results show that *model perplexity* is not directly correlated with generalization abilities, and LSTMs show good results on some phenomena while failing at others. The authors conclude with a call for a wider variety of syntactic phenomena to test on as well as more varied models to be evaluated, in particular with respect to *hyperparameter* selections and randomization seeds.

Another way to assess the network's acquired grammar focuses on a more indirect test of the information encoded in the internal representation, evaluating which aspects of the original syntactic structure can be reconstructed from them through *diagnostic classifiers*. We cite a few examples, representative of different methodologies in this area. Adi et al. (2017), for instance, define auxiliary tasks training classifiers to predict sentence length, word content and word order from vector representations for sentences: using the representation as input for the classifier, they are therefore able to assess the strengths of different embedding methods: interestingly, LSTMs are found to be very effective at predicting word order, among the many proposed tasks.

Perhaps surprisingly, the dominant evaluations do not pay much attention to input, with a few exceptions. In McCoy et al. (2018), for example, the authors specifically address the poverty-of-the-stimulus hypothesis, training an encoder-decoder Recurrent Neural Network to turn declarative sentences into questions through auxiliary inversion. They employ two different training sets, one containing specific cues towards hierarchical generalization through subject - auxiliary agreement, and the other including no specific morphological cues. What they find is that performances are better when agreement is explicitly marked, showing how the recurrent mechanism of the network is apt to exploit the cues in the input to perform the highest possible generalization. In a subsequent study, McCoy et al. (2020) find that only models with an explicit inductive bias (Shen et al., 2018) learn to generalize the *move-main* rule with respect to auxiliary inversion. The statement they draw from their results is that, also for humans, "the hierarchical preference [...] requires making explicit reference to hierarchical structure, and cannot be argued to emerge from more general biases applied to input containing cues to hierarchical structure". Their setup involves however no pre-training task on generic Language Modelling and, in a way, treats the phenomenon of auxiliary

inversion in a vacuum, as if it is the case that children learn question formation as a separate tool from all other linguistic skills.

The relation between the bias shown by language models and the shape of the input has also been specifically mentioned in Hawkins et al. (2020), where the authors examine performances of various pre-trained Neural Language Models, including the LSTM of Gulordava et al. (2018), against a dataset containing human preference judgements on *dative alternations* in various conditions, manipulating the length and definiteness of the recipient argument. The LSTM model showed poorer correlations than the more complex transformers architectures, that seemed to be more sensitive in reproducing human-like biases. The LSTM models were found to be, on the other hand, better at modelling *definiteness* effects. The interesting point of the study is that human intuitions are collected and kept as graded notions, against which the models are tested. And bias is seen as a **proxy** to syntactic abilities rather than as something that is hurting the abstractions process.

Tellingly, in the studies we have mentioned, LSTMs are more or less explicitly treated as psycholinguistic subjects (Futrell et al., 2019). This makes it possible to compare many of the presented studies to adult linguistic performances. But while the practice has yielded meaningful results, the differences between human subjects and models cannot be ignored. The most evident difference concerns *corpora*: the type of text which models encounter in training are often not representative of the variety of contexts an adult speaker has encountered language in, throughout their life. Training is therefore implausible in both *size* (amount of words the model is exposed to) and *genre*: Wikipedia or web-scraped language are widely employed for technical reasons, but these are definitely far from the language speakers experience and produce on a daily basis. While this is a major technical parameter to account for, it also mirrors the idea that the grammar to which speakers are expected to converge is a stable set of rules that in no way depends on the received input. It is the input, instead, that is considered some corrupted evidence of it from which grammar is supposed to be triggered more than abstracted.

2.2.3 *Distributional Semantics in the Usage-Based framework*

Strongly related to the discussion about the role of the input in triggering or shaping linguistic representation is Distributional Semantics (Erk, 2012; Lenci, 2018), a usage-based model of meaning representation that relies on the assumption that meaning in language is an abstraction over the contexts in which linguistic items are used. Distributional Semantics is based on the assumption that semantic relations between lexical items can be approximated by their distribution in linguistic contexts. More specifically, the *distributional hypothesis*

is based on early works by Harris (1954) and basically states that *words that occur in the same contexts tend to have similar meanings*.

A mathematical encoding of the distributional hypothesis can be automatically generated in the form of vectorial representations of linguistic co-occurrences in text. Thanks to their scalar properties, such vectorial representations have shown to be appropriate to approach some of the well-known limitations of formal theories of meaning representation, primarily concerning the relationship between *word meaning* as word ***usage in context*** (e.g., *meaning acquisition, logical metonymy, coercion, idiomaticity*...). As we will see later, the ability to express lexical meaning in a mathematical space allows us to interpret the evolution of linguistic knowledge processed by an LSTM model, so we will provide here additional theoretical background on the approach, foregrounding some of our later claims.

Besides the computational realization in vector space models, *distributionalism* as a general theory of meaning has broader foundations, influenced by Wittgenstein (1953) and, within behavioral psychology, aligned with the theories of Deese (1966), according to which meaning is acquired thanks to the association of co-occurring stimuli. In cognitive sciences, studies such as Rubenstein and Goodenough (1965) showed how similarity judgements and linguistic contexts overlap significantly. Similarly, Miller and Charles (1991) claimed that "words' contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various (linguistic) contexts. Two words are semantically similar to the extent that their contextual representations are similar".

As a framework, Distributional Semantics serves two purposes, both as an **empirical methodology** and as a **cognitive hypothesis** about the nature and emergence of linguistic representations (Lenci, 2008):

1. Empirically, we speak of *weak distributional hypothesis* to refer to Distributional Semantics as a **quantitative method** for semantic analysis: a correlation between semantic content and linguistic distribution is assumed (as in Harris, 1954 distributional procedures). In that sense, the weak distributional hypothesis can coexist within many theoretical linguistics frameworks (e.g., distributional methodologies were employed within the theory of the *Generative Lexicon* by Pustejovsky, 1991).

2. Cognitively speaking, we consider the *strong distributional hypothesis*, which instead deals with the idea that semantic cognitive representations owe their shape and meaning to their distributional properties. Here, contexts have a specific role in the formation of linguistic cognitive representations. This version of the distributional hypothesis pairs very well with usage-based theories of acquisition as a whole, as it grants the input a **causal role** in the process of linguistic knowledge acquisition.

We will make extensive use of the *strong* distributional hypothesis in this thesis, and will show in particular that it gives us a quantitative way to think of graded acquisition as well as speaker variability. But in order to fully appreciate the contribution of the mathematical representation to linguistic theory, we will first expand our literature review to questions of stability and their treatment in linguistic theories.

2.3 STABILITY

Stability and its counterpart, *variability*, relates to different aspects of acquisition, including inter-speaker linguistic differences and intra-speaker behaviour in grammatical judgments. With reference to the continuum we introduced above, we can similarly highlight the stances that we would find at the two ends, however remarking how many possibilities exist in intermediate positions:

- on the one hand, we can see linguistic competence as stable across individuals and over the (almost) entire lifespan of a speaker. This is what a strong version of the *continuity hypothesis* suggests and is what licenses a number of standard procedures in classical linguistics such as heavily relying on the linguist intuitions or on judgements from restricted samples of the population;
- on the other end, we find theories that emphasize how **partial competence** can be observed: this often comes with the idea that speakers present a variety of grammatical intuitions, heavily influenced by context.

In what follows, we will first introduce relevant work in psycholinguistics, before considering two aspects of variability in the analysis of Neural Language Models: behavioural agreement and gradedness in acquisition steps.

2.3.1 *Variability in humans*

Generally speaking, generative approaches have dealt less with acquisitional data than usage-based studies. The ones that do so still minimize the role of variability in children language competences: Crain and Pietroski (2001) for instance state that children soon stabilize their competence reaching to a level that is equivalent to other members of their community, and remark that *humans exhibit mastery of linguistic principles that are not possibly learned*.

The biolinguistic perspective exemplified in Crain, Koring, et al., 2017, and the studies they cite, still heavily rely on the Modularity Hypothesis (Chomsky, 1992) and on the existence of a universal list of parameters (C. Yang and Berwick, 2017), for which however little consensus is reached (Dabrowska, 2015). The idea that children

converge fast and effortlessly to adult grammar is in fact rather controversial in various ways: the very idea of the existence of a unique adult grammar has itself been questioned, as it has been shown that adults have different comprehension abilities depending on a number of factors , and many studies also show that adults and children perform very differently on generalization tasks, both on real language and on artificial input . Moreover, in the generative argumentation, usage-based theories are often reduced as theories advocating for the role of frequency figures as proxies to grammatical generalizations (C. D. Yang, 2004), but the domain-general mechanisms that can be put in use during acquisition go well beyond frequency (Goldberg, 2019).

Despite general age-related tendencies for structure acquisition (Marchetto and Bonatti, 2013, 2015), showing that this is usually mastered by older infants, there are relevant individual differences among children concerning when and how they get to non-adjacent dependency tracking: as suggested by Gomez (2002), Van Heugten and Johnson, 2010 find that the distributional statistics in children's input align with collected behavioural data about remote dependencies. Moreover, Lany and Shoaib (2020) suggest that differences in the ability to track non-adjacent dependencies might be due to the vocabulary size of the subject and Frost et al. (2020) find that word-like and structure-like statistics can be tracked at the same time by children, reinforcing the hypothesis of a strong link between input distributions and the ease of acquisition of more structural phenomena.

Despite the importance of relational reasoning for linguistic proficiency, children seem much less able to detect these kind of analogies than adults, showing a much more varied and unstable set of results, and success seem to depend more heavily on contextual factors. As Goldberg (2019) reports, for instance, the study of G. F. Marcus et al. (1999), where younger children, in a setting where they could not rely on transitional probabilities, recognised instances of an ABA pattern, but that performance is age-related in a non-linguistic setting (G. Marcus et al., 2004). Saffran, Pollak, et al. (2007) showed however that even young children are able to perform the task in a non-linguistic setting, when stimuli are familiar and already categorized. And similarly, Ferguson and Lew-Williams (2016) find that rule-learning abilities are influenced by communicative context, that is thought to enhance learning and generalization of structured input.

On the basis of these constrasting results, Carstensen and Frank (2021) suggest that variation in human relational reasoning is driven by context rather than ability, and that same-different reasoning is a domain-general mechanism and not uniquely human. Moreover, they advocate for the adequacy of graded representations to account for the variability of results, thus drawing a parallelism with new generations of neural network models (Geiger et al., 2022; Santoro et al., 2017).

2.3.2 *Neural Language Models and stability*

Agreement has been one of the main benchmarks for the evaluation of neural models: among the many studies, the one by Arehalli and Linzen (2020) specifically addressed the area of *attraction effects*: despite the subject-verb agreement general rule, real-time human comprehension and production does not always follow the general grammatical constraint, due to a variety of possible syntactic or semantic factors. In the study, six experiments in the agreement attraction literature are replicated using LSTMs as subjects, and the authors find the model to be able to capture the human behavior in three out of the six experiments, showing that generic sequence models without any built-in language-specific mechanism can simulate human grammatical behavior. Moreover, this result is obtained in a setting where the language model is still trained on a specific and semi-formal variety of language (the English Wikipedia) that arguably does not contain as much attraction effects as spoken, informal language. Differently from previously mentioned studies, what makes this attempt particularly interesting is the fact that LSTMs expected to mimic human grammatical behaviour rather than abstract competence, and processing effects are treated as part of the picture.

The relation of LSTMs representations to human sentence processing has also been shown by Hashemzadeh et al. (2020): in a recent study they showed that LSTM-derived representations can be correlated to human brain activity not only for within-distribution language, as was previously known (Hale et al., 2018; Schwartz and Mitchell, 2019), but also for out-of-distribution conditions such as *Jabberwocky* sentences. This is relevant as it places both conditions, i.e., semantically well-formed and Jabberwocky sentences, on the same continuum with respect to the network abilities.

Studies also differ when it comes to their assumptions about the nature of grammar. While part of the literature acknowledges that grammatical constructs may be acquired gradually and with different degrees of completeness, other works posit an '*all-or-nothing*' mechanism, where rules are either fully acquired or simply absent from the speaker's competence, in a way more akin to generative approaches.

Supporting the view of graded acquisition, Wilcox et al. (2018) address the phenomenon of filler-gap dependencies (e.g., the dependence existing between *what* and the "-" mark for a gap in *I know what/*that the lion devoured - at sunrise*). They evaluate the surprisal values assigned by pre-trained language models, namely the Gulordava et al. (2018) one and a much bigger model by Chelba et al. (2013), in two positions: immediately following the gap and summed over the region from the gap to the end of the clause. Their results point out that Neural Language Models (but not *n-gram* models) show high peaks of surprisal in the post-gap position, irrespective of the syntactic po-

sition where the gap happens (either subject, object or prepositional phrase). When considering the whole clause, however, predictions about subjects are much stronger than for the other two positions, in a way that correlates with human online processing results. In a second experiment, they find that the length of the relation correlates with surprisal in the post-gap position, especially for the object case. Overall, their results point to the direction that filler-gap dependencies, and the constrains on them, are acquired by language models, albeit in a graded manner, and in many cases correlating with human judgements on data of the same kind. While both the models they employ were state-of-the-art at the time this article was written, both are trained on rather big amounts of text (one billion words for the Chelba et al., 2013 model) that are implausible as inputs to human speakers.

A different conclusion on a similar task was reached by Chowdhury and Zamparelli (2018): training various versions of Recurrent Neural Networks on the English Wikipedia and a subset of UKWaC (Baroni et al., 2009), they perform the evaluation on a rather controlled set of sentences created by varying the possible subject, main verb etc. within a range. While their results are generally in line with the ones by Wilcox et al. (2018) — i.e., the network is better at dealing with subjects rather than objects — they highlight contrasting results in grammaticality evaluation on affirmative sentences versus Wh-questions (including a gapped dependency), and ultimately state that their model "is sensitive to linguistic processing factors and probably ultimately unable to induce a more abstract notion of grammaticality", thus committing to a strong competence vs. performance distinction.

A similar call for *full abstraction*, as opposed to a graded view of syntactic abilities, is expressed in Marvin and Linzen (2018): a number of English artificial sentence pairs (i.e., a grammatical sentence with its ungrammatical counterpart) is automatically built using a non-recursive context-free grammar. The specific intent is to "minimize the semantic or collocational cues that can be used to identify the grammatical sentence". The dataset explicitly addresses three syntactic phenomena, namely *subject-verb agreement*, *reflexive anaphora* and *negative polarity items*, and two models are evaluated: a simple Recurrent Neural Network language model and a multitask Recurrent Neural Network that solves two tasks at the same time, i.e., Language Modelling and a tagging task that overimposes syntactic information, both trained on a Wikipedia subset. Overall, results are varied both between tasks and, for a single benchmark, between different lexical items: a result that, as the authors say, "would not be expected if its syntactic representations were fully abstract". From a more constructionist perspective, of course, this is perfectly reasonable if we think of abstraction as induced by the association of specific lexical items with grammatical structure and intentions. Moreover, the multitask Recur-

rent Neural Network shows some improvement on specific subtasks, for instance long subject-verb coordination.

Charles Yu et al. (2020a) investigate the grammatical judgments of Neural Language Models (transformers in this case) in a minimal pair setting (i.e., two sentences that differ in their acceptability due to just one grammatical property). The novelty of their study lies in the fact that they focus on how well the model understands the grammatical properties of a particular target noun: the authors find that performances are correlated across tasks and across models. These results suggest two important points: the learning of Neural Language Models seems to be happening in an item-based way, with abstraction emerging from the association with specific lexical items, and that the 'learnability' of an item does not depend on the specific model, but seems to be rather tied to the statistical properties of the input, and those are not restricted to item frequency.

The effects of item frequency on human learning have, on the other hand, been largely and long investigated (Bybee and Hopper, 2001), but it has been equally shown how they are not the most reliable source for predicting productivity (Goldberg, 2019; Gries, 2012). In particular, the relation between frequency and function has received specific attention in the usage-based studies. Meaning plays in fact a key role in accessing and processing individual lexical items (J. S. H. Taylor et al., 2015) and higher-level elements such as multiword expressions (Jolsvai et al., 2020), and is at the very base of the argumentations on productivity that rely on the preemption principle.

2.4 SYSTEMATICITY

We finish this review by considering the issue of *systematicity*, looking at the different ways compositionality and productivity have been treated in the literature. The continuum here, in fact, as we introduced in Chapter 1, is concerned with the amount of predictability we expect from linguistic items:

- in the more compositionality-driven approaches, the meaning of a complex item such as a sentence is built by integrating the meaning of smaller elements such as words, through rules and constraints. In this area, the behavior of higher-order linguistic items is regarded as very predictable and little space is left for indeterminacy or negotiation;

- in productivity-driven approaches, on the other hand, systematicity is intended as the ability, common to all speakers, to use all the available linguistic means to cue the intended meaning (Hernandez et al., 2019). Here *quasi-regularity* or, as we might call it, unpredictability, makes space for linguistic creativity.

2.4.1 *Higher order structures learned through general purpose mechanisms vs. explicit bias*

While the isolation of chunks seems easily tractable through general-purpose mechanisms, one of the major issues that statistical learning models have to face is the existence of non-adjacent structures with very variable aspects on the surface. These kind of long-distance dependencies are common in language, involving verbal structures, as well as higher-order constructions, and might also explain more subtle patterns like agreement throughout the sentence or event-level dependencies: while it is intuitive that we, as speakers, are able to detect this kind of discontinuous patterns, evidence coming primarily from artificial grammar learning is not so strong about it (Gomez, 2002; R. Gómez and Maye, 2005; Newport and Aslin, 2004), it being influenced by a great number of factors such as internal variability and the nature of the elements in the pattern.

In a series of studies, Culbertson, Franck, et al. (2020) and Culbertson, Smolensky, et al. (2012) tested French and Hebrew children on a task involving the creation of noun phrases containing both adjectival and numeral modifiers. Participants, both children and adults, were divided in four groups and exposed to different distributions of nouns modified by pre- or post-nominal adjectives and numerals. Their results show that while adults tend to reproduce or regularize the distributions received in input, children have a strong preference for harmonic orderings, namely those in which both modifiers occur either before the noun or after it. This happens, for children, irrespectively of their first language, as both English, French and Hebrew children were tested, with English being an harmonic language while French and Hebrew show a mixed pre- and post-nominal modification. Despite the fact that Hebrew children will eventually converge to a more complex pattern, the study shows that, independently from their cognitive role, there are general learning tendencies and that human speakers tend to use them unless probably the data is strong enough to override them.

A recent study (Fousheea et al., submitted) track the evolution of *a/the - noun* productivity in English children, comparing it to parental input and to the productivity of the entire *determiner* class. Their work sets the stage for addressing the question in computational work.

2.4.2 *Neural Language Models and sistematicity*

One the side of Neural Language Models, a ground-breaking study involving the evaluation of performances through specialized datasets is the one presented in Gulordava et al., 2018. The study builds on a previous research (Linzen, Dupoux, et al., 2016), that showed how the network acquired *abstract* information about number agreement, albeit

in a supervised setting (i.e., a classifier was explicitly trained to detect that kind of information). Gulordava et al. (2018), on the other hand, train a simple LSTM on a language modelling task and show how this is enough for the network to predict long-distance number agreement, both on semantically sound and nonsensical sentences, concluding that "LM-trained Recurrent Neural Networks can construct abstract grammatical representations". Their model is however trained on a rather consequent amount of data (90M tokens) from a peculiar distribution (a Wikipedia dump). Moreover, they show performances on both semantically sound (e.g., *It presents the case for marriage equality and finds...*) and semantically nonsensical sentences (e.g., *It stays the shuttle for honesty insurance and finds...*), testing different constructions and varying the number of intervening attractors. Their results score well above baselines, showing how Language Modelling encompasses the acquisition of some grammatical knowledge. In their setup and analyses, however, semantics is kept well separated from syntactic abilities, and agreement is evaluated in term of a grammatical rule that the network should be able to apply on top of other choices. During the test phase, the network is fed with the sentence up to the position where agreement happens, and a point is scored if the network attributes higher probability to the lexeme bearing the right agreement than the wrong one.

On this trend, a series of studies showed that networks carrying explicit syntactic bias perform better than vanilla LSTMs. In a recent paper, Lepori et al. (2020) show that a constituency-based network generalizes more robustly than a dependency-based one, and that both outperform a more basic BiLSTM in a classification task: the best results were however obtained on an artificially constructed set of sentences, obtained trough a probabilistic context-free grammar (PCFG) that generates simple transitive sentences (Subject-Verb-Object) with optional modifications introduced through adjectives or prepositional phrases and where all words of a given part of speech are equally likely in all positions. When the BiLSTM is fine-tuned on a distribution of this kind, that explicitly requires the creation of a more abstract representation than lexical co-occurrence, its performances dramatically improve: this suggests that a simple sequential mechanism can be enough if the linguistic signal is structured in a way that abstraction is encouraged.

A different approach is shown in Lakretz et al. (2019): the authors take a rather physiological approach to the assessment of the network's grammatical abilities, investigating how specific neurons specialize in detecting and memorizing syntactic structures. They test the pretrained language model employed in Gulordava et al. (2018) on a number agreement task and on implicit syntactic parsing abilities. By testing the effect of removing each possible neuron from the network, they find two units whose removal reduced the network's perfor-

mance by more than 10%, one responsible for singular long-distance dependencies and one responsible for the plural ones. The drop in performance was however found only on the more complex subtasks, while in the case of simple agreement or with at most two intervening words, the ablation did not have any effect.

One last study to mention is the one carried out by Kuncoro, Dyer, et al. (2018a): differently from previously mentioned studies, a character-based LSTM is employed, therefore lacking explicit word representations. Their results show much lower performances at number agreement with multiple attractors compared to the word-based models. They also test a specific type of networks, Recurrent Neural Network Grammars (RNNGs, Dyer et al., 2016; Kuncoro, Ballesteros, et al., 2017), that estimate the joint probability of string terminals and phrase-structure tree nonterminals, and find them to outperform LSTM language models and syntactic language models without explicit composition functions, highlighting specifically the benefits of top-down parsing as an anticipatory model. Again, performance of simple LSTMs is consistently above random, and their performances are regarded as related to their ability of capturing "patterns that are predictive in most cases", and this is again regarded as a separate ability than the one of acquiring grammar itself.

Giulianelli et al. (2018), using a diagnostic classifier on the internal states of a language model, evaluate representations on number agreement between subject and verb, replicating the study of Gulordava et al. (2018), and use the results of the classifier also to improve the model abilities on the same task: their analyses show that LSTMs might represent subject-verb information agreement at different levels at the same time, retaining both local, short-term information along with a deeper, long-term substrate for successive sequence processing.

The linguistic role of other, non-linguistic biases that Recurrent Neural Networks have has also been under investigation, for instance in Davis and van Schijndel (2020b): in the study the authors examine biases of neural networks for ambiguous relative clause attachments. In a sentence like *Andrew had dinner yesterday with the nephew of the teacher that was divorced*, both *nephew* and *teacher* are available for modifications by the relative clause: from a purely grammatical perspective, both interpretations are equally possible and therefore equally plausible. The authors report however that English speakers have a generic preference for attaching the relative clause to the lower nominal, while other languages such as Spanish show a preference for the higher nominal: while humans are able to overcome their biases and recognize both interpretations, the core of the investigation is not to compare the network's choices to human-elicited responses, but to evaluate similarities in generic linguistic behavior. In a similar setting, investigating the influence of implicit causation on syntactic representations, Davis and van Schijndel (2020a) find the same preference in language

models for agreeing with the lower possible noun, while finding a stronger and more human-like bias in bigger GPT models. The authors find that Recurrent Neural Networks do not resolve ambiguity in a completely human-like way: by comparing an English language model with a Spanish one, differently from human performances, it is shown that both models have a preference for the lower level nominal. The cause for this preference is found in a post-hoc analysis: in Spanish training data (a subset of the Spanish Wikipedia), a distributional bias in favour of low attachment is present, leading the authors to conclude that "standard training data itself may systematically lack aspects of syntax relevant to performing linguistic comprehension tasks". By manually correcting this bias in the input, in a way that an equal proportion of high attachment and low attachment is present, they find that a preference for the higher nominal is learnable by the LSTMs. As the authors themselves claim, this well expresses a general call for better understanding the relationship between training data and the expected performances. Similar tendencies or biases have also been tested on human speakers, as shown by Culbertson, Franck, et al. (2020) and Culbertson, Smolensky, et al. (2012) — cited above.

More systematic evaluations of large models' biases is performed in studies like Warstadt, Zhang, et al. (2020): by pre-training a transformer-based model on increasing quantities of data, the authors find that in order for the models to prefer the more abstract, linguistic generalization over the surface one, not only is a rather large amount of data needed, but data should also explicitly support the linguistic generalization over the surface one.

2.5 DISCUSSION

As we tried to outline, within the computational community the picture is quite variegated: the number of studies aiming at dissecting neural language models' syntactic abilities is increasing rapidly. Investigations are moreover increasingly focusing on the newly introduced, larger transformers model such as BERT (Devlin et al., 2019; Elazar et al., 2020), GPT-2 (Radford et al., 2019b) and the more recent GPT-3 (Brown et al., 2020). The number and variability of employed techniques make it hard to delineate a clear trend: generally speaking, neural language models seem to be quite successful at a number of syntactic tasks, while miserably and somehow unexpectedly failing at others. Such puzzling results point also towards a deeper reflection upon the tasks employed to test the models themselves: Neural Language Models can be seen as dynamic and domain-general learning tools, employing statistical associative mechanisms during the acquisition as well as during the processing and production phases. Many of these features have been similarly highlighted in human linguistic processes, in particular by usage-based theories of grammar acquisition.

As noted by Linzen and Baroni (2020), however, the conclusions drawn from the presented studies largely depend on the idea of competence and performance, lexicon and grammar to which researchers, more or less explicitly, abide by. Specifically, still very few studies explicitly link the performances of Neural Language Models to theoretical works on usage-based formalisms such as construction grammars.

This actually represents a growing area of research: in 2023, a first workshop specifically addressing *Construction Grammars and NLP* is being organized aiming at bridging two fields that are recognized as *complementary, yet currently disparate*.

Construction Grammars and NLP Workshop (2023): https://sites.google.com/view/cxgsnlpworkshop

We mention here some of the studies that investigate Neural Language Models' linguistic abilities while specifically relying on a formalization based on Construction Grammars (an analysis can also be found in Weissweiler, He, et al., 2023). Madabushi et al. (2020) explore to what extent BERT does have access to constructional information. In their study, they design two sets of experiments to assess whether the addition of constructional information affects BERT and how effective BERT is in identifying constructions. They find that BERT retains a substantial amount of information pertaining to semantically specific construction in its internal layers. Weissweiler, Hofmann, et al. (2022) investigate pretrained language models on the English Comparative Correlative construction (i.e., *The X-er, the Y-er*). Their probing tasks consist in checking whether Neural Language Models are able to tell apart sentences that contain the construction from sentences that do not and probe the model's understanding of the construction by asking the model to perform some deductions. The authors report good results for the first task and no significant results for the second. Li et al. (2022) perform an experiment based on Bencini and Goldberg (2000) (mentioned in section above): they find that Neural Language Models prefer sorting by construction rather than by main verb. A further argument in favor of the employment of structural cues by language models is the study carried out by Sinclair et al. (2022), where the authors show how transformer models are sensitive to structural priming and retain hierarchical syntactic information alongside with sequential one.

Besides the evaluation of language models in itself, the uncertainty of results might also suggest a more careful look at Neural Language Models' grammar abilities: most of the mentioned experiments are modeled on psycholinguistic literature, and this entails a number of aspects that, we believe, are often overlooked.

To begin with, most of the studies we mentioned tackle *classic* phenomena, as for instance subject-verb agreement or auxiliary inversion, that are strongly associated to the nativist perspective. In the three continuums that we identified (i.e., *input-as-a-trigger* vs *input with a structural role*, *compositionality* vs *productivity* and *language as expression of the average individual* vs *language as expression of the community*),

these tasks are all placed on the former end: seen as *core* linguistic knowledge, these aspects have a long and important history in linguistic research, but they only partially represent linguistic knowledge when seen from the other end of the continuum. In this direction, a growing set of studies are looking at how **bias affects learning** and influences grammatical performances. While a clear paradigm has not yet emerged, this approach could emphasize the similarities between usage-based approached and data driven Language Modelling.

Many studies also compare human judgements with results from language models. And yet again, whether the network succeeds or fails at reproducing psycholinguistic judgements is often unclear. We believe this mostly stems from the fact that the two sets of data, as they are commonly gathered, are not necessarily comparable. The linguistic background of average European/North American educated adults, which is the most common population for psycholinguistic studies, is incomparably richer with respect to what Neural Language Models are exposed to, which is usually generic text collected from the web. Moreover, data collected from humans is usually task-oriented: depending on the task they are asked to solve, there can be the influence of metalinguistic knowledge or other reasoning skills. For this reasons, the comparison with humans, despite being a fundamental step for the evaluation of Neural Language Models, might not always be fair, as models lack any metalinguistic information and are widely evaluated on pure prediction, which is the basic task that Neural Language Models solve.

Another aspect to be mentioned is that Neural Language Models are compared treated as an **idealized average speaker**, with their predictions being compared to aggregation of human judgements. While this can be regarded as a necessary simplification, it also mirrors the view that there is a universally shared grammar to which speakers, and thus also LMs, converge, and that this convergence is considered as more meaningful. Similarly, LMs are hardly ever compared to developmental data and their training status is often regarded as mature/adult linguistic knowledge. Lastly, most theories of acquisition recognize that comprehension and production abilities are not fully aligned, especially during the first stages. While it is true that the two tasks are more similar in Neural Language Models than in the human case, to our knowledge there is no study specifically assessing differences in Neural Language Models between the two modalities.

In our opinion, this all stems from the idea that *" ... children in the same linguistic community all learn the same grammar"* (see Chomsky, 1965, 1975; Herschensohn, 2009; Lidz and Williams, 2009; Nowak et al., 2001; Smith and Allott, 2016). As Dabrowska (2015) points out, this view continues to be widely espoused, even by cognitive and functional linguists, while in fact considerable differences exist both among children, in their path to language acquisition (Bates et al.,

1988; Dąbrowska, 2004; Richards, 1990) and among adult speakers as well (H. H. Clark, 1997; Farmer et al., 2012; Mulder and Hulstijn, 2011).

This approach also prevents us from investigating **language as a community-level phenomenon**, rather than as an individual one. In usage-based accounts, individuals have a linguistic knowledge which is different from the one possessed by other members of the community, and since they approximate each other's behavior through communication, collective grammars tend to be more systematic than individual grammars (see Dabrowska, 2015; Dąbrowska, 2013; Hurford, 2000).

Our discussion aims at highlighting the fact that a number of **latent biases** are being carried on to computational approaches, maybe lacking full awareness of their downstream effects. What researchers are trying to measure is some notion of *grammatical competence* in Neural Language Models, but, as Lau et al. (2017) argue, this is a theoretical entity not accessible to observation or measurement. With speakers, and not without criticisms (C. Schütze, 2016), the evidence coming from acceptability judgements can count as a measurable phenomenon, but the same principle is not necessarily directly transferable to the Neural Language Models setting.

What we find to be generally missing from the computational state-of-art discussion is the role of the linguist as a *lens* on the gathered data. As Marantz (2005) state, *"one gains the impression from much linguistic writing that grammars in fact are descriptions of data rather than hypotheses about computation and representation"*. Whether distributional generalizations, i.e. a "grammar", which is what LMs are capable of, may or may not extend beyond a specific set of data, is a matter of the hypothesis that the linguist has taken on the generalized categories rather than of the abilities the model itself.

Some words need to be spent also concerning the robustness of intuitions in linguistic argumentation. The attitude of linguists towards speakers' judgements is in fact at times unclear: a great number of the examples cited in literature is in fact ambiguous in terms of validity (Wasow and Arnold, 2005), but they are often presented based on the judgement of few speakers if not the linguist alone. Wasow and Arnold (2005) also argue that usage data tends in general to be overlooked in theoretical linguistics, at least on the more nativist side, where evidence other than intuition is often brought in only as supporting evidence, rather than informing the theory development process.

3

CALAMO: A FORMAL CONSTRUCTIONIST ASSESSMENT OF LANGUAGE MODELS

The content in this chapter partially appeared in: Ludovica Pannitto and Aurélie Herbelot (to appear), "CALaMo: a Constructionist Assessment of Language Models," in Proceedings of the 1st Workshop on Construction Grammars and NLP, *Association for Computational Linguistics,* https://arxiv.org/abs/2302.03589.

As it emerged from the discussion in the previous chapters, it is often the case that the extreme complexity of theoretical tools found in linguistic studies gets cut down by order of magnitudes when it comes to the analysis of language processing using computational modelling. Whether systems acquire any kind of linguistic knowledge remains in fact one of the biggest current research questions in computational modelling. What is meant by *linguistic knowledge* is however often unclear: assumptions that would be clearly stated in theoretical linguistics (e.g. how grammatical abstraction fits into the concept of *language*), are not explicitly discussed by computational studies. Table 3.1 shows for instance some extracts from relevant papers dealing with the analysis of Neural Language Models' linguistic abilities. Irrespective of results proven in the papers, their reference to linguistic knowledge is often without reference to any specific theoretical framework: this makes it hard to determine what the expectations are that researchers have on the model's abilities and what their specific assumptions are.

Most current work also seems to implicitly make a number of **assumptions** about what kind of grammar is supposed to emerge from neural language models, and this underlying choice is often echoed in the most common evaluation settings and in the conclusions that are being drawn from such experiments. Most of these default assumptions are inherited from the nativist Chomskian tradition and the Universal Grammar (UG) framework (Chomsky, 1986; Smith and Allott, 2016), which has pervaded a lot of the computational work on grammar, and continues to do so in the recent literature on neural models. In Chapter 2, we discussed the specific postulates that have been integrated into current frameworks, and whether this integration was warranted, given the architecture and learning behaviour of neural models. We mentioned that ironically, the nativist assumptions that permeate the mainstream computational methodology are at odds with the very nature of the models created by the field, which are essentially based on pattern learning and hugely rely on the **statistical properties** of their training data. In this chapter, we will propose an

Marvin and Linzen, 2019	*We propose to supplement perplexity with a metric that assesses whether the probability distribution defined by the model conforms to the* ***grammar of the language.***
Hu et al., 2020	*Targeted syntactic evaluations have shown that these models also implicitly capture* ***many syntactic generalizations****.*
Linzen and Baroni, 2021	*contemporary DNNs can learn a surprising* ***amount about syntax****, but* ***fall short of human competence****. [...] Implications for the study of* ***human linguistic abilities****.*
Davis and van Schijndel, 2020b	*researchers [...] ask whether those models have learned* ***some linguistic phenomena****.*
Charles Yu et al., 2020b	*Neural language models learn [...] the* ***grammatical properties*** *of natural languages. [...] we focus on the* ***variation in grammatical knowledge****.*
Lakretz et al., 2019	*whether these generic sequence-processing devices are discovering* ***genuine structural properties of language*** *in their training data, or whether their success can be explained by opportunistic* ***surface pattern-based heuristics***

Table 3.1: Some extracts from papers dealing with the analysis of the linguistic abilities Neural Language Models have.

alternative model that instead takes statistical learning seriously and properly integrates theoretical insights with computational methodology. We will call this model **CALaMo**, short for 'Constructionist Assessment of Language Models'. Its goal is to show how to interpret the output of Neural Language Models against a suitable theoretical perspective. In what follows, we will set up the framework as a formal construct before proposing a specific implementation of its various conceptual parts. But before we do so, we will first recap briefly on the main tenets of constructionist approaches, so that our model can be considered against its wider theoretical landscape.

As a theory of acquisition, the notion of pattern learning is almost perfectly in line with the entire family of **constructionist theories**, encompassing construction grammars, cognitive linguistics and usage-based theories at large. Construction grammars themselves are a family of grammatical theories developed starting from the late 1980s: despite being different in their formalizations or in the peculiar aspects they aim at describing, all theories that fall under the constructionist set share some basic principles (Goldberg, 2013; Kay, 1997):

- **Form** and **function** are fundamental traits associated with each grammatical element, and these include the lexicon as well as more complex structures that collect several lexical items along

with empty or partially filled slots. Every kind of structure, lexemes included, is defined as a construction.

- Grammar is made up by the set of **constructions and no other kind of structures**: these are merged in order to produce sentences. No derivational rules are established within the theories. Quoting Goldberg (2006), "It's constructions all the way down".
- All surface linguistic forms are considered equal: there is no distinction between a **core** and a **periphery**, or between productive rules and idiomatic structures. Surface forms are in fact all that matters in the constructionist view.
- The grammar needs to be able to take into account all possible **generalizations** a speaker can produce, based on the available and received data: there remains a difference between the grammar as an abstraction over all the possible generalizations and what the individual speaker actually retains as their linguistic competence.
- Because of the dependence on data, constructions are **language-specific**: where universal tendencies emerge, they are due to functional reasons or general cognitive principles.
- Constructions are organized as a network through inheritance relations in what is often called a **Constructicon**.
- Construction grammars do not impose different (and sequential) modules for processing or composing utterances. Form (including the phonological realization) and function are bound together in the grammatical item stored in the construction.

As cognitive items ("Knowledge of language is knowledge" - Goldberg, 2006), constructions are subject to all known effects of categorization, generalization, prototypicality and so on.

We note that many of these aspects are aligned with the neural approach to language modelling.

- First, language modelling is performed keeping only **surface structure** into consideration.
- Second, **data distribution** is central to the learning trajectory of the model and its final outcome.
- Third, the learned abstractions are **graded**.
- Finally, there is no modularity: information is retained in **a single object** (in the case of neural models, the matrix containing the weights of the neural connections).

That is, while the nativist account and its formalization offered a definition of language and linguistic processes that could be easily implemented through the early computational approaches, their assumptions may not fit contemporary computational procedures in the way that constructionist approaches do.

This chapter starts with an exposition of our formal framework and is set up as the description of an abstract scientific model. The following sections discuss two possible applications of the framework to the fields of language acquisition and language description. In the following chapters, we will showcase these two use case scenarios in real experiments.

3.1 THE CALAMO ALTERNATIVE

In our proposed methodology, CALaMo, we incorporate the usage-based perspective across all three aspects discussed throughout this thesis: *input*, *stability*, and *systematicity*. We will first start with an overview of the way standard approaches deal with those aspects, and briefly state the way that CALaMo offers an alternative. Subsequently, we will formalise our proposal, starting with a definition of acquisition and refining it to integrate the idea of learning as a *process*, and introducing tools to talk about representations and the role of a hypothetical linguist-observer.

In the mainstream evaluation framework, the *input* Neural Language Models are trained on is largely intended as a trigger for linguistic knowledge. This emerges for instance from the approach to prediction that is taken during the evaluation and analysis phase, where input features play little to no role. From a usage-based perspective, instead, the relation between the abstract grammatical structure of the input and the acquired grammar, which then constrains the production of the learner, is strict: construction grammars, differently from the nativist approach, seek to motivate the existence of a particular **form-meaning association** in language. As Goldberg (2006) points out, "functional and historical generalizations count as explanations, but they are not predictive in the strict sense". This has direct impact on the use of test sets, which should be carefully constructed with respect to what the input data is actually able to license as a linguistic generalization. CALaMo differs from standard approaches by considering input data an important factor in determining the shape of the learner's grammatical knowledge: we introduce a methodology to directly compare the distribution of constructions in the input with what is produced by the Neural Language Model in the generative phase.

As far as *stability* is concerned, depending on the view that is taken on the continuity hypothesis, we can see Neural Language Model's grammatical competence either as a binary or as a **gradient**

property. In the first case, we test whether the network is able or not to handle some linguistic phenomenon, while in the second case, as advocated by CALaMo, we are interested in seeing how and why some linguistic aspect becomes more and more salient to the network during training. Our framework was in fact primarily conceived for showing how some linguistic constructions become more relevant during the learning phase of the Neural Language Model: it does not, therefore, presuppose stability as a feature of the language model. Similarly, each trained Language Model is considered as an individual instance. Therefore, no homogeneity across the population is assumed either as it is instead posited in mainstream models: this makes it possible to investigate **inter-speaker variability** or **community-level phenomena**. In our exploratory experiments, we test this with Neural Language Models trained on very small amounts of data: in the mainstream evaluation, framework models are often pre-trained on large corpora that are not standardized and whose effects are often ignored during the evaluation phase (Linzen, 2020).

The **compositionality vs. productivity** perspectives, finally, entail a different organization of linguistic knowledge: the mainstream compositionality perspective tends to set meaning aside, and treat the lexicon as an organized repository of meanings - this is often called *words-and-rules* model or *dictionary-and-grammar* model (Pinker, 1999; J. R. Taylor, 2012): it makes sense, therefore, to test an Neural Language Model's capabilities on semantically nonsensical sentences or to extend the known rules to completely unknown lexical items. In the productivity perspective, instead, meaning is intrinsically part of the process and is treated as a **systematic aspect of grammar**, too. Constructions in CALaMo vary on the two axes of *complexity* (i.e., the number of elements that play a role in the structure) and *schematicity* (i.e., the presence of free variables in the constructions). The two axes vary independently and both influence the productivity of each construction. Each construction is therefore supposed to contain information about its application restriction and, through the relations present in the construction, its **compatibility** with the other linguistic items.

3.1.1 *Formalization: acquiring language*

When talking about Neural Language Models and their linguistic capabilities, the issue of **language acquisition** (A) is often formalized as how much language Λ can be learned by the (artificial) speaker, given a certain level of computational complexity C by being exposed to a certain type of data I:

$$A : C \times I \mapsto \Lambda \tag{3.1}$$

All the components of the equation have been central to the linguistic debate. However, starting from this basic formalization, we identify

two major focus points that we specifically address in our framework. Firstly, the above formula describes acquisition as instantaneous, but it is actually better described as a **process** (Section 3.1.2):

$$A = (a_0, a_1, \cdots a_N) \tag{3.2}$$

From a cognitive perspective the process is fully continuous, while in the artificial scenario, input data is often fed in 'batches'. We can however imagine that, if we had the ability to increase the number of steps at will (i.e., make N larger while keeping constant the amount of data), we could formalize steps small enough to make the two processes comparable.

Secondly, language is often seen as something that the learner has acquired and gained knowledge of. We want to bring back in the framework the role of the **linguist-observer**, that builds an abstraction over the linguistic behavior of the speaker (Section 3.1.3). As the actual knowledge acquired by the speaker is undetectable and only explainable metalinguistically, in a way that is not viable with neural networks (i.e., we cannot ask Neural Language Models what they know about linguistic regularities), we must take into account the fact that we are always analyzing both the linguistic input received by the speaker and the output produced as an effect of the acquisition process through **analytical categories** that are created and used by the linguist-observer. In other words, Λ is not a property of the speaker, but rather a function operated by the linguist-observer. It does not evolve per se during the acquisition process, but rather it helps us detect and characterize the evolution of the speaker's abilities.

3.1.2 *Formalization: acquisition as a process*

All the elements of Equation 3.1 ideally change throughout time as the acquisition process unfolds.

The input I to which the learner is exposed, in a real-life scenario, changes continuously. We can therefore define:

$$I = (\iota_0, \iota_1, \cdots, \iota_N) \tag{3.3}$$

where ι_i is the collection of input data to which the learner has been exposed to in-between a_i and a_{i+1}. Again ideally, with N large enough, each ι_i could even correspond to a single sentence, thus following acquisition in real time.

The computational complexity also co-evolves with the acquisition function, as linguistic knowledge gets incorporated into it. In the human case, the initial state is unobservable, and in the artificial scenario, it is often not interesting as initialization of neural models is random. At step i, instead, the computational mechanism that has

incorporated knowledge up to step $i-1$ is exposed to ι_i. For these reasons, we define:

$$C = (c_\emptyset, c_0, \cdots, c_{N-1}) \tag{3.4}$$

As an effect, Λ identifies different subsets $\lambda_0, \lambda_1, \cdots, \lambda_N$ throughout the acquisition process, namely:

$$\Lambda = \bigcup_{i=0}^{N} \lambda_i \tag{3.5}$$

Each step of the broader process A can be therefore defined as:

$$\begin{cases} a_0 : \iota_0 \times c_\emptyset \mapsto \lambda_0 \\ a_i : \iota_i \times c_{i-1} \mapsto \lambda_i \end{cases} \tag{3.6}$$

3.1.3 *Formalization: how do we observe learned language?*

The notion of language that we introduced incorporates that of **grammar**, namely the analytical categories that we superimpose on the linguistic stream in order to analyze it and its unfolding over time. We do not test language as a cognitive state of the speaker: we intend it instead as a **set of categories** that the observer (i.e., the linguist) considers relevant to the description of the linguistic stream produced by the (artificial) speaker. There exists, therefore, a striking difference between the linguistic stream (either the input perceived or the output produced by the speaker) and its representation through the lens provided by *language* Λ.

If we wanted to be more precise with the notation, we should acknowledge the fact that language, i.e. Λ, as we mean it is actually a **function** by itself, that takes as input some linguistic stream (some observable data) and returns a representation of it. We could therefore rewrite the definition of a_i as

$$a_i : \iota_i \times c_{i-1} \mapsto \Lambda(o_i) \tag{3.7}$$

where o_i is the linguistic stream produced by the speaker as a result of acquisition step a_i.

As we are interested in the categories that are acquired by the speaker and deployed during language comprehension and production, defining $\lambda_{o_i} = \Lambda(o_i)$ allows us to apply the same transformation on the input ι_i to which the speaker is exposed, thus obtaining λ_{ι_i} that is comparable to λ_{o_i} in terms of linguistic categories.

Sticking to the constructionist perspective while trying to make the fewest possible assumptions on the actual content of linguistic knowledge, we hypothesize *language* as made up of a **network of structures** that are supposed to approximate constructions. Since constructions consist of pairs of form and meaning, the concept of grammar includes a semantic dimension that extends beyond lexical level. This can be easily implemented by extending the notion of **vector space models**, that has been extensively explored and used in distributional semantics (Erk, 2012; Lenci, 2008, 2018).

This is a significant departure from nativist perspectives and rhe conventional evaluation framework: meaning is inseparable from grammatical effects, and any model of language acquisition must acknowledge and incorporate its influence in the learning process. If we had to formalize the content of any λ_i, therefore, we could expand it as:

$$\lambda_i = \{(\kappa, \vec{\kappa}) \mid \kappa \text{ is a construction wrt. some linguistic stream}\} \quad (3.8)$$

Unpacking this, we are saying that each obtained construction λ_i is a network of structures. These can be more or less lexicalized, with their schematicity being a proxy for linking the structures in the network as we will explain in the next paragraph. Each construction is associated with a distributional vector (Figure 3.1), which represents its meaning.

3.1.4 *Additional desiderata: the structure induced by* Λ

We defined Λ as a function that takes as input a linguistic stream τ and returns a *construction* λ_τ: a structured repository of form-meaning pairs. In order to describe and explore the internal structure of the construction, we introduce a few auxiliary functions and definitions:

(1): having meaning defined as a distributional space allows for distance computation:

$$d(\kappa_i, \kappa_j) \text{ with } d : \Lambda \times \Lambda \mapsto [0, 1] \quad (3.9)$$

$d(\cdot, \cdot)$ is a metric function that computes the distance between two meaning vectors. Usually

$$d(\kappa_i, \kappa_j) = 1 - cos(\vec{\kappa_i}, \vec{\kappa_j}) \quad (3.10)$$

where $cos(\vec{\kappa_i}, \vec{\kappa_j})$ is the cosine similarity between the two vectors associated to κ_i and κ_j;

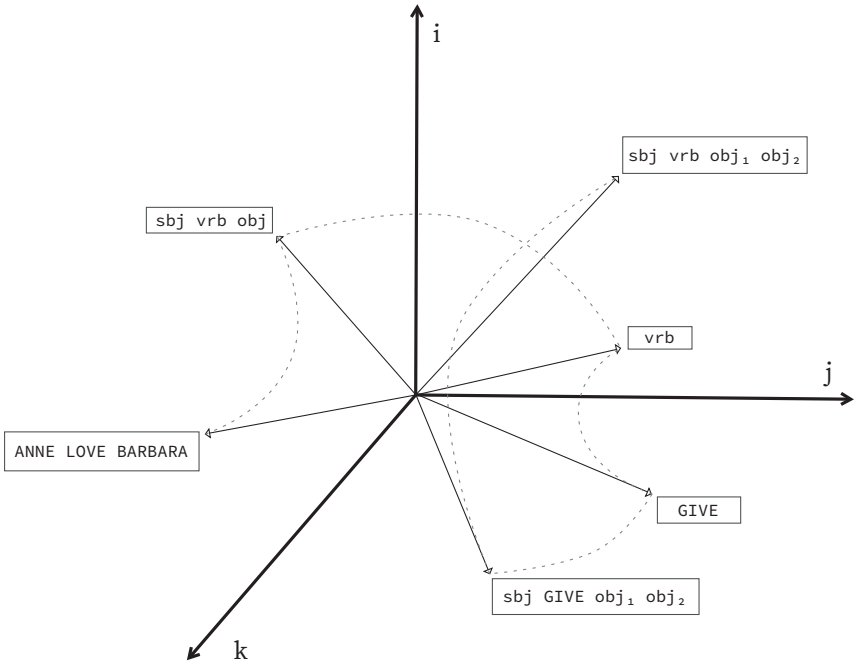


Figure 3.1: The figure shows an example of a Constructicon: it appears as a vector space populated with constructions at different level of *schematicity* (i.e., the fully schematic ditransitive pattern co-exists with fully lexicalized and partially lexicalized constructions). The dotted lines indicate how the Construcricon is organized as a network and constructions can be connected by different types of relations.
As our procedure to build the Constructicon is fully unsupervised, it might contain items that can would be considered as *constructs* rather than constructions from the theoretical perspective. We specifically included the item *ANNE LOVE BARBARA* in the figure to underline this aspect.

(II): constructions bearing different schematicity levels are linked within the network. In order to navigate the network we introduce the function

$$c(\kappa_i, \kappa_j) \text{ with } c : \Lambda \times \Lambda \mapsto \{0, 1\} \tag{3.11}$$

c is a boolean function that computes whether two constructions constitute an *abstraction chain*. For instance, $\kappa_i =$ nsubj, GIVE, iobj, dobj and $\kappa_j =$ nsubj, root, iobj, dobj form a chain with κ_i being a partially lexicalized (hence, less schematic) instance of κ_j.

3.2 CALAMO: USE CASES

In this section, we highlight how the formalization presented in Section 3.1 can be applied to the evaluation of artificial language models. We will focus on two aspects of interest: *acquisition over time* and *speaker diversity*.

3.2.1 *Individual acquisition over time*

The framework can be used to observe how the acquisition process unfolds over time. We can in fact set a number of steps n and observe:

(I) how the **shape** of grammar changes over the course of learning, comparing the various steps, as in:

$$\Lambda(o_1) \sim \Lambda_(o_2) \sim \cdots \sim \Lambda_(o_n) \tag{3.12}$$

(II) how the grammar of the input can be compared to that acquired by the speaker, as in:

$$\Lambda(I) \sim \Lambda(o_n) \tag{3.13}$$

Given a subset $K \subseteq \Lambda(I)$[1] of interesting constructions, we can observe their behaviour over the learning process.

A popular constructionist hypothesis (Goldberg, 2006), for example, states that the meaning of a construction (e.g., the *ditransitive pattern Subj V Obj Obj2*), and therefore its productivity, emerges from the association with specific lexical items in the input received by the learner (e.g., *give* in the case of the *ditransitive*): part of the lexical meaning remains attached to the meaning of the syntactic pattern, and therefore its **distributional properties** with it. Let us assume that the speaker has acquired some construction κ (e.g., the *ditransitive* construction). Once they're able to use it in a productive and creative way (i.e., in more varied contexts than the *give* contexts the construction is strongly associated with in the input), we can use the

1 Actually, we have to make sure that $K \subseteq \Lambda(I) \cap \lambda_0 \cap \lambda_1 \cap \cdots \cap \lambda_n$

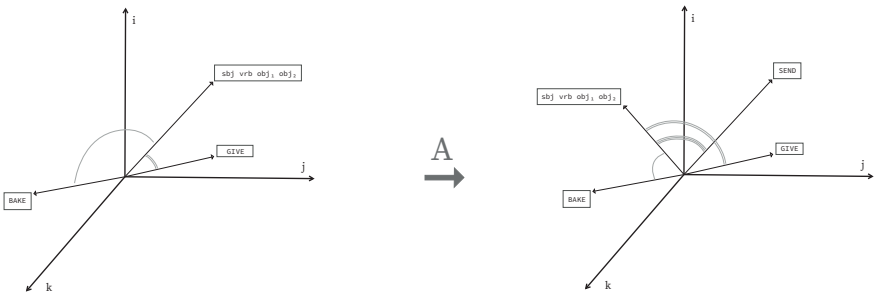


Figure 3.2: The figure gives a visual representation to the hypothesis provided in Goldberg (2006): the meaning of the ditransitive construction is due to its strong association with the lexical item *give* in the early stages of acquisition. Later in life, speakers are able to produce the construction in a wider array of contexts, with the construction itself bringing the *transfer* part of the overall meaning.

proposed framework to check whether the distributional meaning of two constructions $\kappa_i, \kappa_j \in \Lambda(I)$ with $c(\kappa_i, \kappa_j) = 1$ (i.e., with κ_i being a less schematic instance of κ_j) influences the learnability of κ_j as an **independent** construction. We show in Figure 3.2 an example of this process, where the ditransitive vector moves away from the *give* vector in the course of acquisition.

The notion of *abstraction chain* introduced before helps us test this hypothesis, as we can check the behaviour of the chain (κ_i, κ_j) at each timestep. We can denote $\kappa_i^{\lambda_k}$ the construction $\kappa_i \in \lambda_k$ and similarly $\kappa_j^{\lambda_k}$ the construction $\kappa_j \in \lambda_k$, through distributional analysis we can capture how the contexts in which κ_i and κ_j vary, and whether this variation is associated with grammatical generalization. We expect, in fact, $d(\kappa_i, \kappa_j)$ to increase during acquisition:

$$d(\kappa_i^{\lambda_a}, \kappa_j^{\lambda_a}) \leq d(\kappa_i^{\lambda_b}, \kappa_j^{\lambda_b}) \;\; \forall\, a, b \mid a \leq b \tag{3.14}$$

If κ_j is produced in contexts that do not perfectly overlap with those where κ_i is produced, this indicates that the speaker has gained a productive use of construction κ_j, which is recognized as an independent construction from κ_i. If, conversely, their distance decreases during acquisition, we might deduce that the speaker has recognized κ_j as unnecessary by restricting its application cases to those of κ_i.

3.2.2 *Language as the expression of a population of speakers*

We are often interested in defining *grammar* in terms of what can be considered **shared linguistic knowledge** among the speakers. A core aspect of construction grammar is in fact conceiving language primarily as a social and external phenomenon, as opposed to nativist theories that focus on its inner nature. By means of the framework, we

can analyze grammar as an abstraction over the linguistic productions of a population of P speakers

$$\Pi = (\sigma_1, \sigma_2, \cdots, \sigma_P) \tag{3.15}$$

We can define the grammatical conventions deployed by the community Π as

$$\Lambda_\Pi = (\lambda_{\sigma_1}, \lambda_{\sigma_2}, \cdots, \lambda_{\sigma_P}) \tag{3.16}$$

This allows for modelling variation between the acquisition process of the different speakers. Speaker σ_i might be exposed to a unique series of input material

$$I_{\sigma_i} = \iota_0^{\sigma_i}, \cdots, \iota_N^{\sigma_i} \tag{3.17}$$

that does not necessarily coincide with that of speaker σ_j.

In this setting, we can for instance investigate what is learned *no-matter-the-input*, and what is instead specific or idiosyncratic for each speaker. Let us define:

$$X(\kappa_i, \sigma_j) = \begin{cases} 1 & \text{if } \kappa \in \Lambda_{\sigma_j} \\ 0 & \text{otherwise} \end{cases} \tag{3.18}$$

as a counting function. X evaluates to 1 if the construction κ appears in the production of speaker σ_j and 0 otherwise. This way we count how many speakers use construction κ productively. By means of X, we can define:

$$G_{\geq p} = \left\{ \kappa \mid \sum_{i=0}^{P} X(\kappa, \sigma_i) \geq p \right\} \tag{3.19}$$

as the set of constructions that we can observe in the linguistic productions of p or more speakers.

G_P would for instance be the set of constructions shared by all speakers in a population, and could be therefore identified as the set of *core* constructions in Λ^Π. When, instead, $p \ll P$, we are observing constructions that are not shared by a significant amount of speakers in the population, and their use can therefore depend on specific input instances or tendencies in subgroups of speakers. Following the same logic we can of course also just define $G_{(\sigma_i, \sigma_j)}$ as the constructions that are common to the two speakers σ_i and σ_j.

By means of G, we can define $\widetilde{\Lambda_G}$ as an approximation of the function Λ, which only uses the categories that are retained in G. $\widetilde{\Lambda_{G_{\geq P}}}$ would for instance be a function that considers only linguistic knowledge shared by the entire population Π, while $\widetilde{\Lambda_{\sigma_i}}$ would be

restricted to the construction λ_{σ_i}. Considering speakers σ_i and σ_j, with their respective produced linguistic outputs O_{σ_i} and O_{σ_j}, we can produce and compare $\widetilde{\Lambda_{G_{\sigma_i}}}(O_{\sigma_j})$ and $\widetilde{\Lambda_{G_{\sigma_j}}}(O_{\sigma_i})$: respectively, what speaker σ_i is able to retrieve from O_{σ_j} and what speaker σ_j is able to retrieve from O_{σ_i}.

The fact that speakers use the same constructions κ to build their linguistic productions does not of course ensure that the corresponding meanings $\vec{\kappa}$ coincide.[2] Different speakers, depending on the input they have been exposed to, and to the partial randomness attributed to computational mechanisms, could associate different meaning spaces to the same construction. Given two speakers σ_i and σ_j, and a sentence s, we can therefore compare the portions of λ_{σ_i} and λ_{σ_j} meaning spaces that are activated to linguistically (de)compose the sentence s.

3.3 CALAMO IN PRACTICE

We now turn to the implementation of our formal model. Our architecture is composed of three main blocks, shown in Figure 3.3 and Figure 3.4, each playing a fundamental role in the acquisition process:

1. The **input** I available to the learner. Neural language models are often trained on unrealistic input and more in general the specific features of the input are overlooked. In the specific case of child-directed language, as we already mentioned in the previous chapters, it has been shown how relevant the distribution of the input appears to be, in order to rightly interpret the child's results. Here, we specifically describe some examples of **child-directed** data that we employed in the experiments described in the next chapters.

2. The **neural learner** itself, with its computational complexity C. Our framework is based on the evaluation of the learner's output in a free generation task. (That is, we make the neural model '*babble*' and record its utterances for further analysis.) Since we rely on usage-based approaches, the idea is that we want to **infer** the linguistic knowledge of the learner based on surface patterns, just as we do in the case of human subjects, for which it is not feasible nor useful to use other and more invasive approaches. In our experiments we employ vanilla LSTM networks, which will be briefly described here.

3. The **linguistic representation** Λ, which instantiates the constructions. This is superimposed on both the input and the learner's produced output and its built thanks to the linguistic categories provided by theory. The kind of **patterns** we are able to extract of course depends on the linguistic categories that we abide by,

2 This makes sure that G_{σ_i} does not coincide with λ_{σ_i}

making the structure and content of the construction theory-dependent. Our final representation is encoded in a distributional semantics framework, taken not only as a formalism but also as a usage-based approach to derive the meaning of linguistic items from their occurrences in the linguistic environment. We chose Universal Dependencies as a formalisms to represent language: from dependency-parsed data, we extracted constructions in the form of *catenae*, encoding them as distributional vectors for meaning representation.

The rest of this chapter describes each of these blocks in turn, together with the implementation choices we made. We justify the suitability of our design decisions with respect to our overall framework.

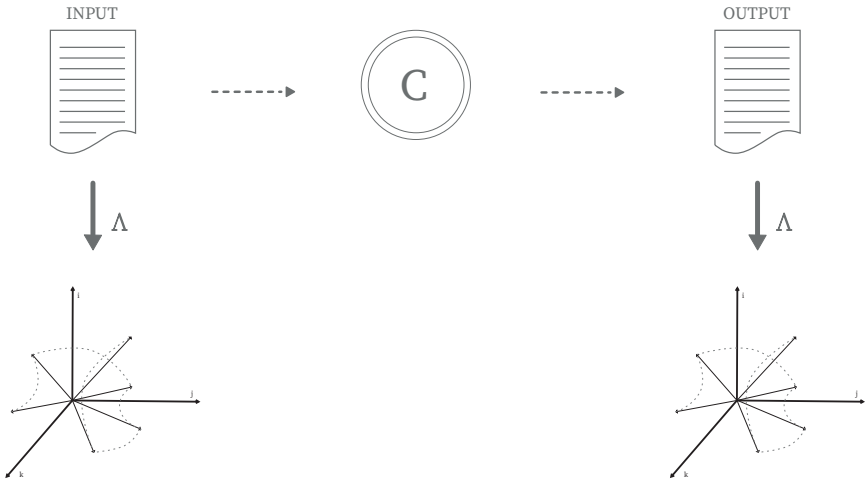


Figure 3.3: The figure shows the main components to the acquisition process, as modeled here: the *input*, costituted by raw text, a central *acquisition mechanisms* (denoted by the letter *C*) and and *output* generated after acquisition in the form of raw text. Constructions can be build, by means of Λ, from both *input* and *output* data.

3.3.1 *Input*

Because of the traditional sharp distinction between *competence* and *performance*, the role of the input and the linguistic environment has been minimized by theories in the realm of Universal Grammar. Usage-based theories, on the other hand, have granted the input a central role to the end of explaining why language is structured as it is (Christiansen and Chater, 2016a; C. J. Fillmore, 1988; Goldberg, 2019; Hoffmann et al., 2013; Kay and C. J. Fillmore, 1999): one of the striking points to make here is that in the usage-based framework, the acquisition problem is framed as an incremental process. Acquiring language

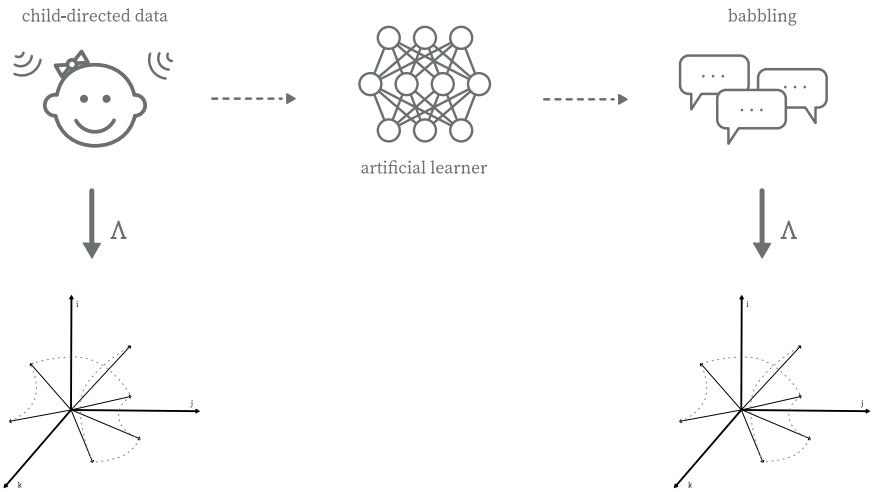


Figure 3.4: The figure shows how the components illustrated in Figure 3.3 are interpreted in CALaMo. In particular, the *input* is *child-directed*, the *acquisition mechanism* is instantiated by a *general purpose vanilla LSTM* and the *output* is constituted by raw test generated (i.e., *babbled*) by the network.

essentially entails **learning how to process** the linguistic input in an error-driven procedure, where full linguistic creativity and productivity are acquired gradually by speakers (Bannard et al., 2009), building up on knowledge about specific items and restricted abstractions.

In this sense, the specific features of the language on which Neural Language Models are trained cannot be overlooked when it comes to describing their acquired grammatical abilities. Compared to what a child is exposed to during the most crucial months of language acquisition, Neural Language Models are trained on an input that is often unrealistic in size: the LSTM introduced in Gulordava et al. (2018) is for example exposed to 90M tokens, and sees them multiple times over training. It is hard to come up with a precise estimate of the amount of language children are exposed to during the years of acquisition, as the variation depends on a huge number of factors including the socio-economic environment (Bee et al., 1969) or the societal organization (Cristia et al., 2019). Hart and Risley (1995), in a seminal work, estimate that, by the age of 3, welfare children have heard about 10 millions words while the average working-class child has heard around 30 millions. Finally, the domain of the data also matters: child-directed language is characterized by **specific features** (Matthews and Bannard, 2010) that are not present in the most widely used corpora (specifically, those that contain data harvested from the Web such as Wikipedia or UKWaC). On the basis of such considerations, we advocate for feeding Neural Language Models with text which is as close to child-directed speech as possible.

In defense of a child-motivated input

Artificial models are often trained on language that is in many ways dissimilar from the language that a child is exposed to in the period of acquisition, at least in most cultures. While language is generally skewed, child-directed language specifically presents features which, following the constructivist hypothesis, enable abstraction of grammatical categories.

Skewedness influences the performance of artificial learners too. Wei et al. (2021) for instance examine the influence of both absolute and relative verb frequency in a *subject-verb* agreement task on BERT (Devlin et al., 2019). The model seems to perform both full grammatical abstractions and item-based learning, depending on the **frequency** of the learned item, but information on sufficiently infrequent lexemes do not get abstracted to grammatical patterns. In order to examine the effect of skewedness, the authors manipulate the input by modifying exactly one variable between absolute and relative frequency of the target verb: what they conclude is that BERT is heavily influenced by the skewed training distribution.

The frequency distribution of the input is however not the only factor to take into account. The genre of the text also matters. Language sources such as Wikipedia or UkWaC (Baroni et al., 2009), which are widely used when training Neural Language Models, may also miss or **underrepresent** some key linguistic constructions (i.e., *open questions*) whose acquisition is considered crucial in non-nativist approaches.

For these reasons, we considered data from a child-motivated perspective in our experiments: we collected portions of existing corpora, with specific attention given to developmental language (Section 3.3.1).

Datasets

As sources of data, we explored the following resources:

CHILDES - Child-directed utterances of the North American (NA) and British (UK) portions of the CHILDES database (MacWhinney, 2000);

GUTENBERG - Books and newspapers from 18 children-related bookshelves of Project Gutenberg (incl. literature, instructional books and others);

OPENSUBTITLES - Movie and TV series subtitles from the OpenSubtitle corpus (Tiedemann, 2012), filtered on the content-rating label;

SIMPLEWIKIPEDIA - A 2019 snapshot of Simple English Wikipedia, written in basic and learning English.

All the considered resources are in the English language. We describe them in more detail below.

The CHILDES Corpus derives from the Child Language Data Exchange System (i.e., CHILDES) Project (MacWhinney, 2000), which was developed with three main aims:

CHILDES Project: https://childes.talkbank.org/

(I) automate the process of data analysis;

(II) obtain consistent and fully-documented transcriptions;

(III) augment the quantity, in terms of language spoken and age range, of available child-related data.

Most of the corpus is composed of **spontaneous interactions** between young monolingual children and their parents or siblings.

The Gutenberg Project contains over 60,000 **books**, mostly older works for which U.S. copyright has expired. The collections are organized in *Bookshelves* that aggregate books of related content.

We focused on the *Children's Bookshelf* (Table 3.2).

Gutenberg Project: https://www.gutenberg.org/, the resource is currently unavailable in Italy since May 2020 in compliance with a decree of the court of Rome over copyright violation.

Childrens Myths Fairy Tales etc.	Childrens Book Series
Childrens Verse	Harpers Young People
Childrens Instructional Books	Little Folks
Childrens History	Childrens Picture Books
School Stories	Childrens Biography
Childs Own Book of Great Musicians	Childrens Literature
Golden Days for Boys and Girls	The Nursery
Childrens Anthologies	Childrens Religion
St. Nicholas Magazine for Boys and Girls	Childrens Fiction

Table 3.2: Categories present in Project Gutenberg, in the *Children Bookshelves* section.

The OpenSubtitles (OPUS) Corpus (Lison and Tiedemann, 2016a; Tiedemann, 2012) is a collection of **movie subtitles** derived from the OpenSubtitles project.

OpenSubtitles https://opus.nlpl.eu/OpenSubtitles-v2018.php, http://www.opensubtitles.org/

The corpus is wide and provides multilingual resources. In order to abide by the child-motivated principle, we semi-automatically selected the entries related to child-suitable content by checking the parental guidance rating (Table 3.3) associated to the movie title on the internet movie database (IMDb), when available. We note here that parental guidance does not mean that the show or movie was explicitly produced for children (i.e., who produced the movie not necessarily planned it to be *child-directed*). However, the content was defined as *suitable* for a children audience.

IMDb: https://www.imdb.com/

Simple Wikipedia is a collection of **Wikipedia projects**: the articles contain more basic lexicon and shorter sentences and are generally tailored to children and adults learning English.

Simple Wikipedia: https://simple.wikipedia.org/wiki/Main_Page

Television	
TV-Y	appropriate for all children
TV-Y7	appropriate for children age 7 and above
TV-G	suitable for all ages
Movies	
G	all ages admitted. Nothing that would offend parents for viewing by children.

Table 3.3: Description of Parental Guidance labels used to select material to include in the subcorpus.

stats	**childes**	**opensub**	**simplewiki**	**gutenberg**
$\|\|C\|\|$	14.1M	7.6M	23.4M	85.6M
$\|\|V\|\|$	53K	119K	451.8K	338.4K
#sentences	3M	1M	1.5M	4.2M
$\|\|C\|\|$/#sent	4.7	7.6	15.6	20.3
TTR	.0038	.016	.019	.004
HTR [:6M]	.002	.006	.014	.005

Table 3.4: Features of the child-motivated corpora considered in this thesis.

The Simple English Wikipedia, as of February 2022, has a little over 200k articles and is the 49th largest Wikipedia available.

Processing

UDPipe: https://ufal.mff.cuni.cz/udpipe

All the corpora were parsed with `udpipe`: from each corpus, we built by random sampling a train-dev-test set, which we further analyzed through Profiling UD toolkit (Brunato et al., 2020), in order to get a better understanding of the differences existing in **syntactic complexity** among the sources.

Considerations about the distribution

While all the considered corpora are child-motivated, they present quite different features. Some of them are presented in Table 3.4: although the Table contains very basic information such as Type-Token Ratio (TTR) or average sentence length, we can already notice how, for instance, speech-derived language (CHILDES and Opensubtitles) distinguished itself from written language (Gutenberg and OpenSubtitles) as far as average sentence length is concerned.

We then performed a slightly more refined analysis on same-size sampled of the three main resources (CHILDES, Opensubtitles and Simplewikipedia). Some figures about what are traditionally consid-

ered linguistic complexity parameters are reported in Table 3.5. Many differences can be spotted, we just cite a few:

(I) *average sentence length* is much higher for simplewikipedia, reflecting the existing difference between written and spoken language;

(II) *hapax-token ratio*, namely the number of hapaxed divided by the total length of the text, is one order of magnitude higher for simplewikipedia, capturing the fact that the resource spans over a much larger array of topics than the other two;

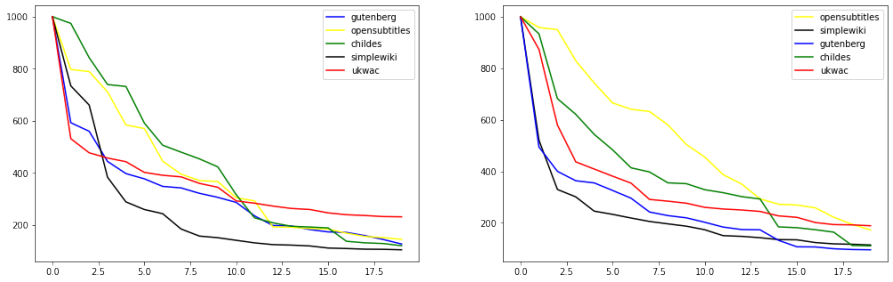
(III) CHILDES shows the lowest *average -arity of verbal roots*: on average, in CHILDES, verbal roots have fewer dependants than in the other resources;

(IV) finally, *pronouns* and *wh-words* seem to be completely missing from simplewikipedia.

		CHILDES	opensubtitles	simplewikipedia
	number of sentences	527k	430k	184k
	number of tokens	2,799k	2,421k	2,570k
	vocabulary size	21k	45k	114k
	average sentence length	5.311	5.634	13.965
	average word length	3.403	3.722	4.494
	type-token ratio	0.007	0.018	0.044
	hapax-token ratio	0.003	0.008	0.025
	lexical density	0.543	0.552	0.600
	average depth of sentences	2.779	2.891	4.386
	average *arity* of verbal roots	2.831	4.176	4.948
PoS of root	**Verb**	60.25%	51.18%	59.87%
	Noun	20.90%	22.35%	33.25%
	Adjective	6.13%	9.75%	4.58%
	Pronoun	2.36%	1.16%	0.05%
	Wh-word	3.60%	2.26%	0.08%

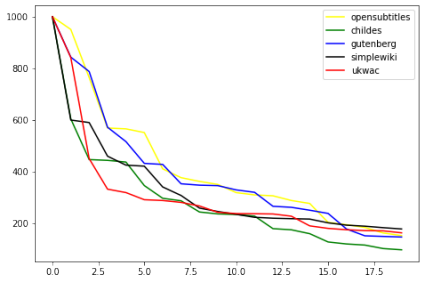
Table 3.5: The table shows some simple figures that highlight some of the differences among the resources we gathered.

We then considered the distribution of vocabulary items bearing some specific grammatical features and checked whether their distribution would be significantly different in different corpora. If we followed constructionist approaches, in fact, we would expect distributions of lexical items for some grammatical aspects to be more skewed in child-directed data than in normal text: this stronger association is what allows for picking up the feature as a construction. We specifically show distributions for *-arity of verbs* (Figure 3.5), *form* (Figure 3.7)

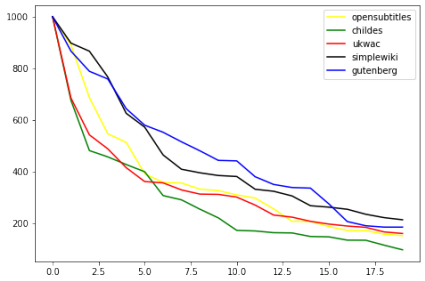


(a) *-arity* vale 1

(b) *-arity* vale 2



(c) *-arity* vale 3



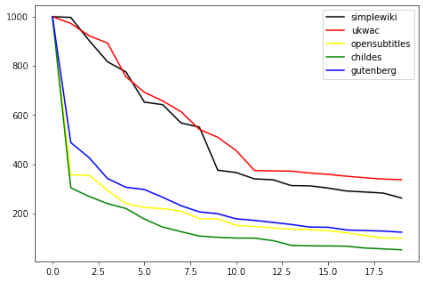
(d) *-arity* vale 4

Figure 3.5: The plots show the distribution of lexical items appearing in the corpora with different *-arities*.

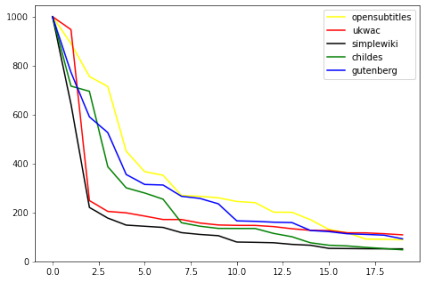
and *tense* (Figure 3.6) of verbs and *pre- and post- verbal (nominal) subjects* (Figure 3.8). We also included *UkWaC* in the analysis as it is among the most widely employed resources for training Neural Language Models. For better comparing the distributions, the plots show just the first 20 positions of the zipfian curve and all distributions are represented considering 1000 as the highest frequency value.

3.3.2 *The neural learner: Long Short-Term Memory networks*

Recurrent Neural Networks (RNNs), and more specifically the 'Long Short-Term Memory network' or **LSTM** (Hochreiter and Schmidhuber,

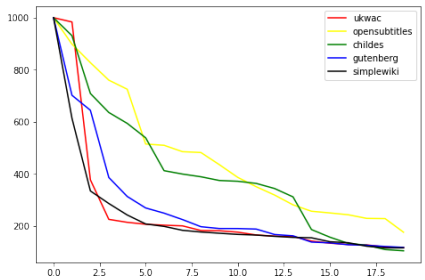


(a) Past tense

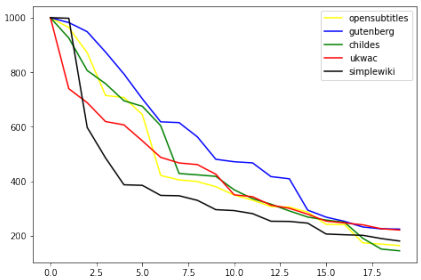


(b) Present tense

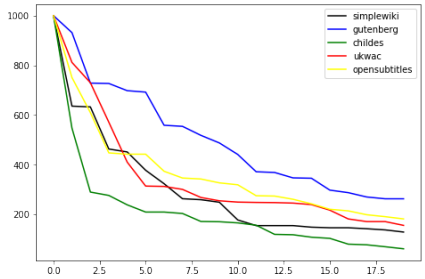
Figure 3.6: The plots show the distribution of lexical items appearing in the corpora with different *verb tenses*.



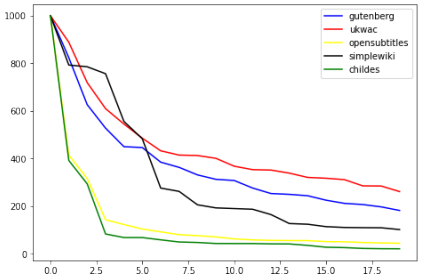
(a) Finite verb form



(b) Infinite verb form

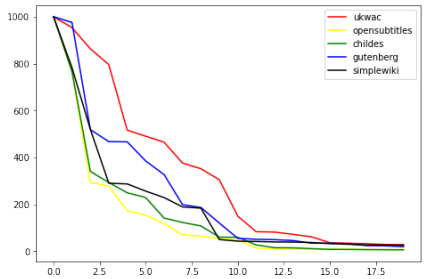


(c) Gerund verb form

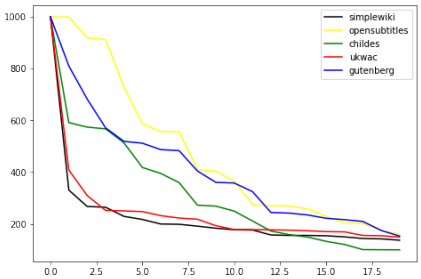


(d) Participle verb form

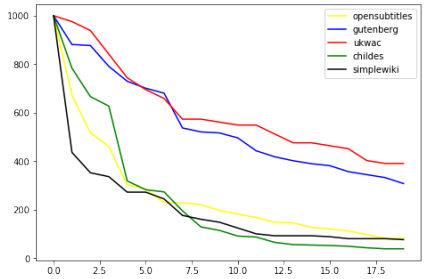
Figure 3.7: The plots show the distribution of lexical items appearing in the corpora with different *verb forms*.



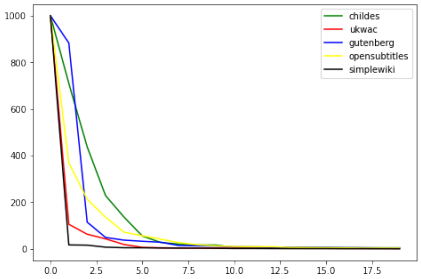
(a) presubjnoun caption



(b) presubjverb caption



(c) postsubjnoun caption



(d) postsubjverb caption

Figure 3.8: The plots show the distribution of lexical items appearing in subject position and verb position for pre- and post- verbal nominal subjects

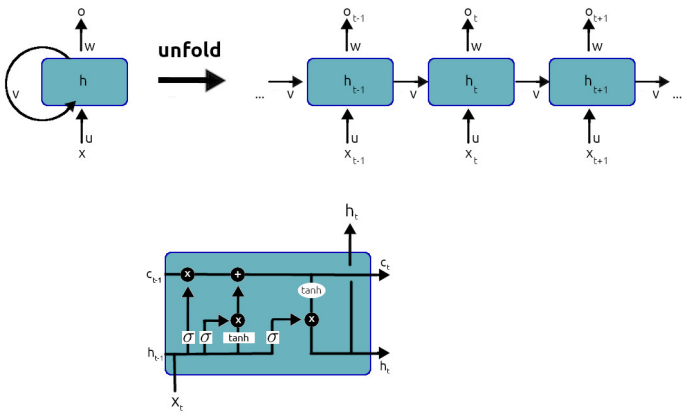


Figure 3.9: The basic structure of an LSTM. The top part of the figure shows the general architecture of any Recurrent Neural Network, involving recurrence at the level of the hidden layer. The bottom part shows the specific structure of an LSTM, involving three different gates (represented as σ).

1997), are among the most common architectures and the ones with the longest history in Language Modeling. The success of LSTMs stems from their ability to keep track of **long-term dependencies**, using an internal structure involving different type of '*gates*', as described below.

Figure 3.9 gives a high-level overview of an LSTM's architecture. As shown in the upper part of the figure, all Recurrent Neural Networks present a chain-like structure: at each time step t, the network's output is computed based on both the input of time t (x_t) and the network's state at time $t-1$ (h_{t-1}). Often, Recurrent Neural Networks are described as a simple multilayer perceptron with a recurrence in the hidden layer. By '*unfolding*' the recurrence, we end up with multiple copies of the same network connected to each other by learnable weights..

A complete and easy to read guide to LSTMs can be found at https://colah.github.io/posts/2015-08-Understanding-LSTMs/

When zooming into the LSTM cell (lower part of Figure 3.9), we additionally find that it has the ability to regulate how the two kinds of information (input and previous state) are weighted towards the computation of the output. The first gate, the *forget gate*, evaluates C_{t-1} (a representation of the previous state different from h_{t-1}) against x_t and learns what information to keep from previous step, including it in a vector f_t. Next, a candidate value for the current state $\hat{C}_t$ is computed along with the *input gate* vector i_t that weights how much of the input will contribute to the current state. Finally, the state of the cell C_t is computed by weighting C_{t-1} with the forget gate vector f_t and the $\hat{C}_t$ with the input vector i_t. h_t is then computed from C_t.

Since **contextual information is maintained** from one prediction step to the next, the output of the network at time t depends on a subset of the inputs fed to the network across a time window.

The LSTM learns to regulate its **attention** over this time window, deciding what to *remember* and what to *forget* in the input. It has been shown that this mechanism allows the network to deal with long-distance dependencies, for instance gender agreement across a piece of discourse, in a much better way than is provided by vanilla Recurrent Neural Network architectures.

In the context of this thesis, LSTMs are a useful framework to investigate learning in a purely predictive setting – as opposed to cases involving an innately biased model. A purely predictive LSTM, trained over surface input, will learn to generalize over the sequences it has seen in the training data to perform a next-word (or next-character) prediction task. Whether it actually encodes some aspects of grammar in this process is subject to debate, and the question has generated a substantial amount of work in Computational Linguistics, as we saw in Chapter 2. A biased model, such as the Recurrent Neural Network Grammars (RNNGs) (Dyer et al., 2016; Kuncoro, Ballesteros, et al., 2017), will on the other hand enforce the incremental generation of a parse tree at the same time as the word prediction task. At each step, the prediction of the model is conditioned on both the previous word sequence and the current state of the parse tree. Expectedly, LSTMs that carry explicit syntactic bias and specifically highlight the benefits of top-down parsing as an anticipatory model (Kuncoro, Dyer, et al., 2018b) tend to perform better in experiments. But the question asked by usage-based theories is to what extent such hard-coded biases could be learned from language exposure only. So in our experiment, we will use the most basic form of LSTM, namely a character-based model trained on sequence prediction without any additional bias.

The consequence of our architectural choice is that we can regard our LSTM model as a **simulation** of a language learner, equipped with only a basic prediction mechanism. Following the practice of usage-based approaches, we would like to infer the linguistic knowledge of the learner from the patterns we observe in its speech. To achieve this, we expose the model to some amount of text, as described in Section 3.3.1, to simulate acquisition. We then use the trained model to generate the same amount of tokens as it was originally exposed to, and perform an analysis of the features encoded in the produced language.

As in experiments involving human participants, all we have access to is the '*flat*' output of the model, i.e. the surface representation of the generated sentences. In order to retrieve patterns in the underlying structure, we first have to convert the produced data into a more hierarchical representation, which can be chosen to match any theory of interest. We describe next how to perform this step.

3.3.3 *Linguistic Representation*

As we have mentioned earlier, we conceive Λ as a function of the linguist-observer, who **builds** an abstraction over the linguistic behavior of the speaker: in fact, we deem the actual knowledge acquired by the speaker to be undetectable and only explainable metalinguistically. In other words, we do not consider Λ to evolve per se during the acquisition process, but rather we regard it as a tool to quantify and characterize the evolution of the speaker's abilities. Each λ_i constitutes then a *construction* relative to a specific set of linguistic data, built through the categories available in Λ.

Two choices appear necessary at this stage:

- whether or not to annotate linguistic categories on raw data and what representation framework to choose
- how to extract constructions from data and consequently build the construction

Concerning the former point, our ultimate aim is to evaluate the abilities of Neural Language Models: we are therefore evaluating their linguistic productions against the child-motivated input described above. Both kinds of data (child-directed speech and the network's babbling), for different reasons, are often ill-formed and need to be automatically processed in order to build the constructions. We therefore chose to parse the data following the universal dependencies (Nivre et al., 2020) formalism.

We should note here that constituency-based representations have been prevalent in the description of natural language syntax, becoming primarily associated with derivational theories. Due to the Fregean view of compositionality, they have also become the natural building blocks for meaning composition. Dependency representations have, on the other hand, re-gained popularity over constituency representations in the last decades, showing desirable properties from a computational perspective: they are in fact suitable for representing a wider array of languages, and by means of dependency representations it is easier to partially represent ill-formed sentences. Moreover, their output is often used as a basis for semantic graphs. Generally speaking, they take a more functional approach to language description, in line with cognition oriented-approaches.

As Osborne (2006) rightly points out, however, the notion of constituent as *any node plus all the nodes that that node dominates* is possible in a dependency framework as well, and actually the predictions of a dependency formalism about constituents correlate better with standard constituency tests such as *topicalization* or *cleft* (Osborne, 2018). We will follow this insight and introduce next the Universal Dependencies framework, and the specific way in which we extract a construction out of the representation.

Universal Dependencies

Universal Dependencies (UD, Nivre et al., 2020) are a cross-linguistic annotation framework developed on the basis of *Stanford dependencies* (de Marneffe, Dozat, et al., 2014; de Marneffe, MacCartney, et al., 2006; de Marneffe and Manning, 2008), *Google universal part-of-speech tags* (Petrov et al., 2012), and the *Interset interlingua* for morphosyntactic tagsets (Zeman, 2008).

The project was first released in 2014 with the general aim of providing a universal **inventory of categories** and annotation guidelines to enable consistent annotation of similar structures across languages, while still allowing for language-specificity (Figure 3.10).

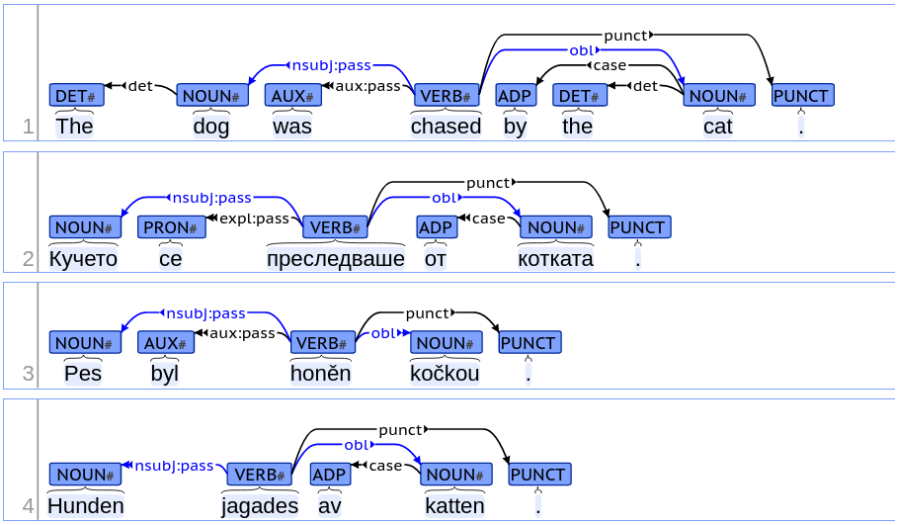


Figure 3.10: The same sentence in four different languages, as represented in the Universal Dependencies framework. The examples are taken from `https://universaldependencies.org/introduction.html`.

The project has now reached version 2, with nearly 200 treebanks available in over 100 languages. A great number of tools have been developed and released to work with UD, including processing pipelines that provide automatic annotation of data.

Cross-lingual alignment is mainly obtained by basing the annotation on **content words** rather than function words, following the so called *lexicalist hypothesis*: this privileges semantically contentful relations and aligns well with the usage-based hypothesis.

The construction

In constructionist theories, grammar is essentially composed of a structured network of constructions, each with its associated meaning. In order to explain our choice for the representation of constructions, let us first examine what definitions of 'construction' have been given in some fundamental constructionist works.

We recall here three definitions in particular:

C. J. Fillmore, 1988 - by 'grammatical construction' we mean any syntactic pattern which is assigned one or more conventional functions in a language, together with whatever is linguistically conventionalized about its contribution to the meaning or the use of structures containing it

Goldberg, 1995 - C is a construction $\iff$ C is a form meaning pair (F, S) such that some aspects of F or some aspects of S is not strictly predictable from C's component parts or from other previously established constructions

Goldberg, 2006 - learned pairings of form with semantic or discourse function

All definitions stress the presence of **meaning** (or *semantic function*, as in the last one): in particular, in Fillmore's definition (C. J. Fillmore, 1988), meaning is described as *the contribution to the meaning or the use of structures containing it*. This is particularly in line with our interpretation of the meaning of a construction, as we will detail in "Distributional Vector Space Model" (paragraph 3.3.3) below.

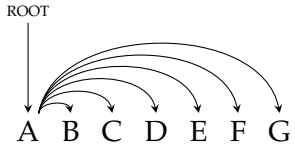
The second definition, provided in Goldberg (1995) highlights instead the **unpredictability** of some aspects of either the form or the meaning of the construction, in order for a pattern to be considered as such. We implement this by introducing a weighting process on the patterns that we extract, considering part of the construction only those that exhibit sufficient statistical association among their components.

Lastly, C. J. Fillmore (1988) suggests that a construction can be, in principle, *any grammatical pattern*: we therefore chose Catenae (Osborne, Putnam, et al., 2012) as a basis to fill up the construction.

an exception should be made for the formalisms derived from the FrameNet project (https://framenet.icsi.berkeley.edu/

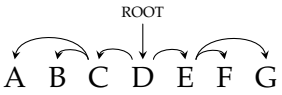
Generally speaking, constructionist approaches seem to lack a shared representational framework, relying on box diagrams or Attribute-Value Matrices to describe the traits of the fragments they study. The structures introduced by Osborne (2006) are characterized instead as fundamental meaning-bearing units (Osborne and Groß, 2012), in line with the theoretical tenets of Construction Grammars, thus being ideal candidates for the lexicon (or *'construction'*) postulated in such theories: *catenae* have in fact been applied in the description of construction-like structures (Dunn, 2017; Osborne and Groß, 2012) and allow for the representation of non-adjacent structures while encompassing the notion of constituent as well (Osborne, 2006, 2018).

CATENAE A catena is defined as "*a word, or a combination of words which is continuous with respect to dominance*" (Osborne, Putnam, et al., 2012): given a dependency tree, this definition selects a broader set of elements than the definition of constituent, which can be seen as a subtype of catena as "A catena that consists of a word plus all the



(a) The case of a flat structure, where all nodes are linked to the *root*: from a tree like this we can extract $2^6 - 1$ catenae, each one containing A plus a subset of its children nodes.

(b) The case where nodes are arranged in a full dependency chain: here the number of catenae corresponds to the number of substrings that could be extracted from the linear signal, that is 20.



(c) The case of a hierarchical structure, typically what we would find in linguistic trees, where the counts are less trivial to make. In particular, for each node we find that the number of catenae *rooted* in that node can be estimated depending on the number of catenae *rooted* in his children nodes, and depends therefore on the specific structure of the tree.

Figure 3.11

words that that word dominates". Unlike constituents, catenae can include both contiguous and non contiguous words. They however capture something more refined than generic subsets of sentence items, as the elements are grouped depending on the syntactic links holding in the sentence.

From a graph-theory perspective, catenae form subtrees (i.e., subsets of nodes and edges that constitute a tree themselves) of the original tree.

Let us consider, for example, the structures represented in Figure 3.11a, Figure 3.11b and Figure 3.11c: the same elements (nodes A to G) are arranged differently in the structure of dependency tree, and this leads to a different number and composition of catenae.

As a concrete example, Figure 3.12 represents a dependency tree, and Table 3.6 the structures that can be extracted from it: considering

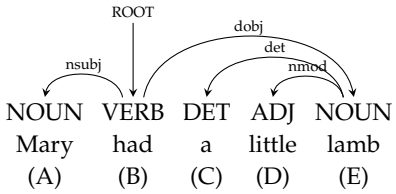


Figure 3.12: The dependency representation of the sentence *Mary had a little lamb*, annotated with morpho-syntactic and syntactic information.

Strings	*A, AB, ABC, ... B, BC, BC, ... E*
Catenae	*A, B, C, D, E, AB, ABCE, ABDE, ABCDE, ABE, BCE,BDE, BE, CE, DE, CDE*
Constituents	*A, ABCDE, C, D, CDE*

Table 3.6: Possible structures that can be extracted from the dependency tree in Figure 3.12

the lexical level, we can extract *Mary had lamb*, *had a lamb*, *a little lamb* as catenae. As the morpho-syntactic and syntactic levels are available, however, we can also extract partially filled structures as *Mary had NOUN*, *nsubj VERB dobj* and so on.

Of interest for our analysis, Construction Grammars argue that grammar items above the lexical level bear meaning themselves, and that this emerges from patterns of usage. According to Goldberg (2006), for example, the meaning of the ditransitive pattern *Sbj V Obj Obj2*, and thus its productivity, emerges from its strong association with *give* in child-directed speech: part of the meaning of *give* remains attached to the construction. A natural, and promising (Rambelli et al., 2019), solution to represent the semantics of catenae is given by Distributional Semantics (Harris, 1954), where each element of the *'construction'* is implicitly described in terms of its context of use (Erk, 2012; Lenci, 2018). We will see in Section 4.3 how we can use such distributional representations to investigate the level of abstraction of our network's babbling.

Catenae were fist formalized in O'Grady (1998) (previously also in Hudson, 1984) in relation to the syntax of idioms. Idioms come with various levels of schematicity, ranging from fully specified chunks (i.e., *red herring*) to discontinuous structures that seem to elude any possible definition in terms on constituents (i.e., some allow an open genitive position such as *the cat got X's tongue*, others allow non-idiomatic modifiers such as *jump on the [MOD]* bandwagon*).

This issue of recognising and employing units that are discontinuous on the surface level, as constructions often are, represents a crucial aspect of language learning.
In O'Grady (1998)'s terms, idioms are therefore subject to the *continuity constraint*, namely the fact that its components must be part of a syntactic chain, for which he gives the following definition: *The string x...y...z... (order irrelevant) forms a chain iff x dominates y and z or x dominates y and y dominates z* — the original definition in O'Grady (1998) uses the term *licenses* instead of *dominates*, that is instead introduced in Osborne (2006). The same principle is soon generalized to constructions, of which idioms constitute a subclass, and that could be characterized as chains at more abstract level of representation as syntactic categories or semantic classes of overtly filled items. Allowing schematic categories in the picture naturally removes the dividing line between fully lexicalized items and partially filled constructions

that however show systematicity and idiomaticity like *V one's heart out*, where the noun-particle combination is fully idiomatic while the verb slot shows few constraints and contributes to the overall meaning in a less idiomatic and more compositional fashion (i.e. *sing your heart out* vs. *work your heart out*).

DISTRIBUTIONAL VECTOR SPACE MODEL Each catena in the construction is paired with a distributional vector that represents its meaning in a vector space model.

Distributional Semantic Models (DSMs) can be generally represented as tuples $< T, C, W, S >$ where:

- T are the target elements (i.e., the items for which the Distributional Semantic Model provides a representation);
- C are the linguistic contexts with which T co-occur;
- W is a context weighting function (or the objective function in the case of predictive models);
- S is a similarity measure between the produced representations.

In our setting, constructions are our target items and we consider sentences as the context window in which they may co-occur: namely, we consider two constructions as co-occurring if they both appear in the same sentence. This approach collapses both paradigmatic and syntagmatic relations together: a co-occurrence is considered both when an item occurs within a catena (for instance, a lexical item within a free slot in a catena) and when two catenae simply co-occur in the same sentence with no overlap of any sort.

We base our distributional model on classic matrix- or count-models, which generalize the basic idea developed in Information Retrieval by Salton et al. (1975): each grammatical item is assigned a n-dimensional count vector (with n being the number of considered catenae). Raw frequencies are then weighted with Positive Pointwise Mutual Information (PPMI) and the matrix is projected into a smaller space through Singular Value Decomposition (SVD, Klema and Laub, 1980) technique. Reduced, implicit vectors help mitigate the issue of data sparsity.

$$ppmi(x,y) = max(log_2 \frac{p(x,y)}{p(x)*p(y)}, 0)$$

4

ACQUISITION OVER TIME

The content in this chapter partially appeared in: Ludovica Pannitto and Aurélie Herbelot (2020), "Recurrent babbling: evaluating the acquisition of grammar from limited input data," in Proceedings of the 24th Conference on Computational Natural Language Learning, *Association for Computational Linguistics, Online, pp. 165-176, https://www.aclweb.org/anthology/2020.conll-1.13.*

Neural Language Models have consistently demonstrated great capabilities in reproducing natural language surface patterns. But although they show good performances when tested on very specific grammatical abilities (Gulordava et al., 2018; Lakretz et al., 2019), as we outlined in Chapter 2, it remains unclear how and to what extent grammatical abilities emerge in artificial language models, and how this knowledge is encoded in their representations. From a theoretical point of view, the principles upon which Neural Language Models are trained and their results seem to contradict some of the most well-known tenets of nativist theories, such as *poverty of the stimulus* (Chomksy, 1959; Chomsky, 1968), while the framework in which they are evaluated remains biased towards the traditional nativist principles.

As we have seen in Chapter 2, a considerable amount of literature has investigated the ability of Neural Language Models to acquire *grammar*, and the analysis of Artificial Neural Network-based language models is by no means a recent endeavour (Lewis and J. L. Elman, 2001; McClelland, 1992). Lately, the general tendency has been to analyze the **inner-workings** of networks, and the specific type of knowledge they acquire (Alishahi, Chrupała, et al., 2019; Linzen and Baroni, 2020). But a clear trend has not yet emerged (Linzen and Baroni, 2020), and we advocate this is an effect of the **latent bias** present in the evaluation framework, rather than due to the Neural Language Models' abilities themselves: Neural Language Models are generally expected to reach *systematicity* through *unbounded* compositional generalization, while usage-based approaches have an interest in how Neural Language Models reproduce *quasi-regularities* rather than full algebraic compositionality.

Through the lens of CALaMo, we depart from the standard account by testing the grammatical abilities of a Neural Language Model in a usage-based perspective. Specifically, we are interested in the following questions:

(1): what kind of **structures** are abstracted and reproduced by the network, depending on the specific input stream received during training;

(II): how the **abstraction** of *schematic patterns* takes place over time.

4.1 FRAMEWORK: CALAMO FOR ACQUISITION

In CALaMo we introduce a methodology to evaluate the acquisition of structures during the training of an individual (artificial) speaker.

We set up two different paradigms for addressing points (i) and (ii) above, which will be more specifically described in Section 4.1.4. Before detailing them, however, we will describe how CALaMo adapts to this scenario (Section 4.1.1), the data and models employed (Section 4.1.2) and the full experimental pipeline (Section 4.1.3).

4.1.1 *Formalization: more details*

Our aim is to test how much grammatical structure can be **induced** from linguistic input through a pattern-finding mechanism such as that provided by Artificial Neural Networks. Therefore, we fix the level of computational complexity to a vanilla, *character-based* LSTM, which we train exploring different sources of input in a specific range $\{I_i\}$, selected based on their complexity level (Section 4.1.2).

We then use the trained models to generate some amount of text (which we call *'babbling'*), comparable in size to what the model has received as input.

We finally explore the construction λ built out of it by the model.

In Chapter 3, we defined the acquisition process as a function A which operates on input I and an acquisition mechanism with its computational complexity C, and returns some linguistic representation $\Lambda(O)$. By means of this formalization, we can therefore write:

$$A(\text{LSTM}, I_i) \rightarrow \Lambda(O_i) \tag{4.1}$$

meaning that here we operate acquisition A by feeding an LSTM with some specific input I_i in order to obtain the output O_i which is then represented by means of Λ.

4.1.2 *Data and Neural Language Model*

As mentioned in Chapter 3, our corpus is composed of three parts, each presenting different features with respect to linguistic complexity:

- Child-directed utterances of the publicly available North American and United Kingdom portions of the CHILDES database (MacWhinney, 2000);

- movie and TV series subtitles from the OpenSubtitle corpus (Lison and Tiedemann, 2016b), filtered by content-rating label (G for movies and TV-Y, TV-Y7, TV-G for TV series), available from *The Movie Database* and the *Internet Movie Database*;
- a 2019 snapshot of Simple English Wikipedia, an English-language edition of Wikipedia written in basic English.

The Movie Database: https://www.themoviedb.org/

We used the three different subcorpora for question (i) and only the first part (i.e., the CHILDES portion) for question (ii).

The portions vary in size: for our experiments we randomly and with uniform probability (Listing 4.1 extract sentences from each source so that the total number of tokens approximates 3 millions (10% are kept for validation and 10% for testing).

```
def reservoir_tokens_number(corpus_sentences, number_of_tokens):
  reservoir = []
  len_sampled = []
  considered_tokens = 0
  sentence_number = -1

  for sentence_num, sentence in enumerate(corpus_sentences):
    considered_tokens += len(sentence)
  if considered_tokens < size*1.2:
    reservoir.append(sentence_number)
    len_sampled.append(len(sentence))
  else:
    j = random.randrange(sentence_num)
    if j < len(reservoir):
      reservoir[j] = sentence_num
      len_sampled[j] = len(sentence)

  x = 0
  i = 0
  while x < size and i < len(reservoir):
    x += len_sampled[i]
  i = i+1
  return list(sorted(reservoir[:i]))
```

Listing 4.1: Reservoir sampling (Vitter, 1985) pseudocode

For each of the considered corpora, a character-based LSTM is trained on the tokenized, raw text. To do so, we slightly modify the *PyTorch* (Paszke et al., 2019) implementation of a vanilla LSTM, adapting it to a character-based setting: as we have to compare language from different sources with different genres, employing a character-based setting allows us to keep the vocabulary size constant and moreover reduce assumptions on vocabulary distribution. We run a *Bayesian optimization* process (Nogueira, 2014–) to select the best hyperparameters for the corpus.

Vanilla LSTM in PyTorch: https://github.com/pytorch/examples/tree/master/word_language_model

4.1.3 *Experimental pipeline*

For each subcorpus, a Neural Language Model is trained through a two-step Bayesian optimization. The hyperparameters' regions where optimization was performed are reported in Table 4.1. Given the optimal hyperparameters, reported in Table 4.2, we then consider an *acquisition step* to be a sequence of 5 epochs of training.

	step 1	**step 2**
batch size	20-129	20-40
emsize	40-401	350-500
hidden	40-401	350-500
nlayers	2,4	3,4
dropout	0-0.5	0-0.2
learning rate	0.001-1	0.8-1
epochs	50-150	30-70
seq length	10-101	15-50

Table 4.1: Hyperparameters' regions used for step 1 and step 2 of the Bayesian optimization procedure.

corpus	**target**	**batch**	**emsize**	**hidden**	**nlayers**	**dropout**	**lr**	**epochs**	**bptt**
CHILDES	1.063	28	371	495	3	0.112	0.96	37	39
OPENSUBTITLES	1.148	20	353	495	3	0.163	0.92	46	48
SIMPLEWIKI	1.171	28	371	495	3	0.112	0.96	37	39

Table 4.2: Target values (perplexity) and parameters of the best models selected by the Bayesian optimizer.

We therefore train a model for each subcorpus with the parameters of Table 4.2. After each *acquisition step*, we also make the model generate some output. More specifically, we sample utterances until we reach approximately the size of the input. Since the network is character-based, but the remainder of the extraction process (namely, the linguistic Artificial Neural Networkotation and the consequent extraction of catenae) is based on sentences (i.e., we extract catenae from each sentence independently), we have to take this aspect into account when sampling, as we need to end up with full sentences while still sampling on a character-by-chatacter basis.

Therefore, we fix a number of iterations (150), and sample a starting letter at the beginning of each one, based on the distribution of letters at the beginning of sentences in the original corpus. For each iteration, we sample a variable number of sentences in order to match the number of sentences in the input (359 for OpenSubtitles, 159 for Simple Wikipedia, and 597 for CHILDES). We also set a maximum number of characters per sentence (as the average sentence length in the input,

plus two standard deviations). This requirement is necessary, as the model does not always manage to reach an *end-of-sentence* character and we want to avoid infinite character sampling, especially in models from earlier epochs, which may not yet fully have acquired the notion of a complete sentence.

The text is then processed with the `english-ewt` model from *Universal Dependencies 2.3 Models* (Straka and Straková, 2018), in order to get dependency representations and extract catenae from there.

english-ewt model: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2898

We extract catenae both from the input corpus and from each *babbling* stage. To do so, we perform a *recursive depth-first visit* of dependency trees (Listing 4.2). That is, if the node A is a leaf, then the only possible catena is the one containing A itself; otherwise, all catenae rooted in A are formed by A plus a (eventually empty) combination of catenae rooted in its children nodes.

```
def recursive_C(A, tree_children, th=5):
  # if A is a leaf
  if A not in tree_children:
    return [[A]], [[A]]
  else:
    found_catenae = []
    list_of_indep_catenae = [[[A]]]
    for a_child in tree_children[A]:
      c, all_c = recursive_C(a_child, tree_children)
      found_catenae += all_c
      list_of_indep_catenae.append([[None]] + c)

    X = []
    for tup in itertools.product(*list_of_indep_catenae):
      new_catena = list(sorted(filter(
          lambda x: x is not None, sum(tup, []))))
      if len(new_catena) <= th:
        X.append(new_catena)

  return X, X+found_catenae

def extract(sentence):
  children = {}
  tokens = {}
  postags = {}
  rels = {}
  excluded_relations = ["discourse", "fixed", "flat", "comound",
                        "list", "parataxis", "orphan", "goeswith"
                            ,
                        "reparandum", "punct", "dep"]
  for token in sentence:
    position, word, lemma, pos, _, morph, head, rel, _, _ = token
    if not pos == "PUNCT" and not rel in excluded_relations:
      if head not in children:
        children[head] = []
      children[head].append(position)
```

```
      tokens[position] = word
      postags[position] = "_"+pos
      rels[position] = "@"+rel

  if 0 in children:
    root = children[0][0]
    _, catenae = recursive_C(root, children)

  for catena in catenae:

    tokensandpostags = [[tokens[x] for x in catena],
                        [postags[x] for x in catena],
                        [rels[x] for x in catena]]

    temp = [(0, 1, 2)] * len(catena)
    X = list(itertools.product(*temp))

    for c in X:
      cat = []
      for i, el in enumerate(c):
        cat.append(tokensandpostags[el][i])
      cat = tuple(cat)
      if len(cat) > 1:
        yield cat
```

Listing 4.2: Pseudocode for catenae extraction procedure

With this procedure, we extract catenae from sentences with length between 1 and 25. For efficiency reasons, we exclude catenae longer than 5 elements. Many structures are thus extracted, not all of which happen to be significant in our framework: we in fact want to populate the construction with patterns that are associated in a statistically significant way. As catenae as pieces of the lexicogrammar, in fact, frequency is not the only relevant parameter and elements (i.e., the different components of a catena that can be either lexical items, parts of speech or syntactic relations) should be positively associated in order to be recorded as objects.

We therefore weigh the produced structures with a *multivariate* version of **Mutual Information** (MI), based on Van de Cruys (2011):

$$MI(x_1;...;x_n) = f(x_1,...,x_n)\log_2 \frac{p(x_1,...,x_n)}{\prod_{i=1}^{n} p(x_i)} \tag{4.2}$$

where

$$p(x_1,...,x_n) = \frac{f(x_1,...,x_n)}{\sum_{(x'_1,...,x'_n)} f(x'_1,...,x'_n)} \tag{4.3}$$

In the equation, $x_1, \cdots, x_n$ are the components of a catena, $f(x_1, \cdots, x_n)$ is the frequency of the catena and p indicates probability.

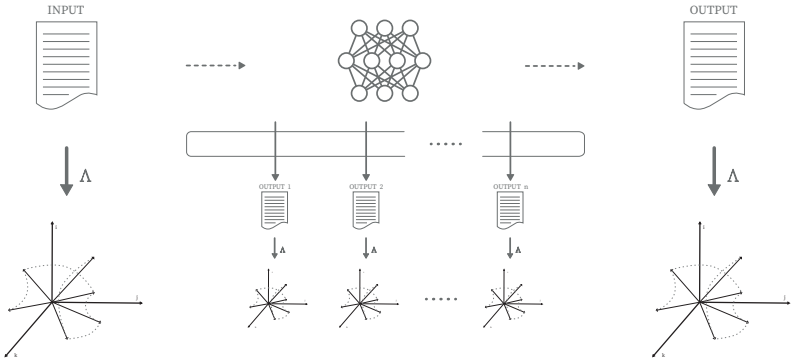


Figure 4.1: The figure shows how the CALaMo framework is adapted to the acquisition setting. Instead of producing constructicons just for the *input* data and the generated *output*, the model is frozen each *k* epochs of training (5 in our case). At each acquisition step some data is *babbled* by the network and from that data a constructicon is built, with the aim of representing linguistic knowledge at that specific stage of acquisition.

It is important to remark that the linguistic Artificial Neural Networkotation process (except for the *tokenization* step) and the catenae extraction processes are completely independent from the language modeling performed by the LSTM, which is only fed with raw text and is therefore completely agnostic about the linguistic categories superimposed by the parser. This is important specifically with respect to the idea, introduced in Chapter 3, that the categories that we are evaluating are not intrinsically part of the observed data, but they are rather abstractions performed by the linguist-observer.

At the end of this process, we obtain seven intermediate constructions for CHILDES (λ_1^C to λ_7^C), nine for OpenSubtitles (λ_1^O to λ_9^O) and seven for SimpleWikipedia (λ_1^S to λ_7^S). The difference is due to the different number of *acquisition steps*. The number of epochs on which each model is trained is in fact different (see Table 4.2): as each step is constituted by five training iterations, this results in a different number of *acquisition steps* among models.

Finally, the distributional space for each λ_i^X is obtained on the basis of counting co-occurrences between constructions within the same sentence. That is, for each construction, at each step of acquisition, we have a vector space showing the relative positions of different constructions. We will use these spaces to show the dynamicity of the acquisition process, i.e. how each individual construction moves across the vector basis as a result of new exposure to linguistic input.

4.1.4 *Formalization of research questions*

We can now turn back to our **question (i)**, namely, what kind of structures are abstracted and reproduced by the network, depending on the specific input stream received during training.

In order to address this question, we evaluate:

- the correlation of pairs of constructions built from input data:

$$\lambda_I^C \sim \lambda_I^O \sim \lambda_I^S \tag{4.4}$$

- the correlation of pairs of constructions built from subsequent steps in training:

$$\forall X : \lambda_i^X \sim \cdots \sim \lambda_n^X \tag{4.5}$$

- the correlation between the input construction and constructions built at each step i:

$$\forall i, X : \lambda_I^X \sim \lambda_i^X \tag{4.6}$$

- the correlation between the input construction and constructions built at each step i from a different input:

$$\forall i, X, Y : \lambda_I^X \sim \lambda_i^Y \tag{4.7}$$

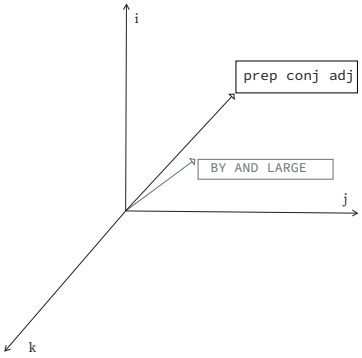
Question (ii) asked how the abstraction of schematic patterns takes place over time. To answer this question, we restrict ourselves to the CHILDES subcorpus: this choice derived from the fact that the three corpora exhibit different distributions, more specifically CHILDES is the only proper child-directed one, while Opensubtitles can be more properly defined as child-suitable and simplewikipedia, while being purposedly written with simplified language, shows distribution which is very distant from that of spoken language.

We consider abstraction chains (κ_i, κ_j) in I (i.e., in Λ(CHILDES)) and compute

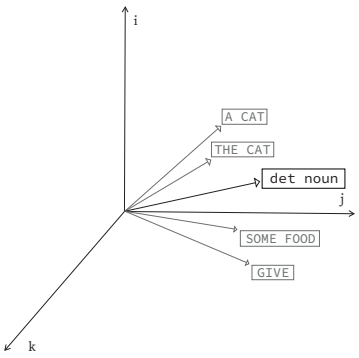
$$d(\kappa_i^{\lambda_7}, \kappa_j^{\lambda_7}) - d(\kappa_i^{\lambda_1}, \kappa_j^{\lambda_1}) \tag{4.8}$$

for each abstraction chain, namely the difference in cosine similarity between *acquisition step* 7 and *acquisition step* 1. We refer to this value as *distributional shift*.
Similarly we compute

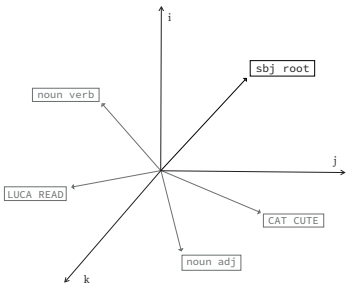
$$d(\kappa_i^{\lambda_I}, \kappa_j^{\lambda_I}) \tag{4.9}$$



(a) Constructions with high similarity values.



(b) Constructions with intermediate similarity values.



(c) Constructions with low similarity values.

Figure 4.2

where I refers to the input data: this is simply the distributional similarity of the two constructions in the input space.

We then group all pairs by κ_i and κ_j and compute the *average distributional shift* for each κ_i to its more schematic instances and for each κ_j to its instantiations. Both for κ_i and κ_j, three bins are considered, based on *average distributional shift*.

Our hypothesis is that catenae that undergo the highest shifts during training are those showing intermediate levels of similarities in the input distributional space. Indeed, pairs with very high input similarities are unlikely to exhibit abstraction: according to constructionist intuition, their distributional similarity means that the catena that is part of the *construction* is the least abstract one, and there is no need for the more abstract category. Low similarity pairs, on the other hand, may simply contain unrelated catenae (Figure 4.2a, Figure 4.2b, Figure 4.2c).

4.2 STRUCTURES THAT ARE ABSTRACTED AND REPRODUCED BY THE NETWORK

Examples of *babblings* are reported in Table 4.3 and Table 4.4.

Our first analysis demonstrates that the language generated by the LSTM **reproduces** the distribution of the input, and that this happens well beyond the lexical level: in other words, the network has acquired **statistical regularities** at the level of grammatical patterns, and is able to use them productively to generate novel language fragments that adhere to the same distribution as the input.

Figure 4.3 shows the extent of this approximation for various pairs: it emerges from the plot that correlations are very high *within* each corpus (on average, 0.935 for CHILDES, 0.929 for OpenSubtitles and 0.917 for Simple Wikipedia). In particular, the correlations between the best models (BM) and the respective input series (I) show values that are among the highest, demonstrating that the network acquires structures and reproduces them with a distribution that almost perfectly matches the input. On the other hand, it is clear that different corpora show different distributions, as correlations between pairs of input series I and best models show much lower values.

The complete set of correlation values is reported in Table 4.5, Table 4.6, Table 4.7 and Table 4.8:

Overall, CHILDES scores the best correlation values, probably due to the specific features of child-directed speech, specifically its repetitiousness (E. V. Clark, 2009). OpenSubtitles interestingly shows intermediate properties, sharing quite a lot of catenae with CHILDES,[1] while Simple Wikipedia shows a completely different distribution.

A selection of the most and least associated catenae for each subcorpora can be found in Table 4.9, Table 4.10 and Table 4.11.

4.3 ABSTRACTION OF SCHEMATIC PATTERNS DURING LEARNING

Our second analysis relies on the idea that we can state that the network *has learned some grammar* once it is able to use an acquired pattern in a productive and creative way. Following the basic hypothesis of CxG, stated in (ref to section), we expect this generalization ability to evolve during training and the distributional properties of patterns to be in relation with the grammatical abilities of the network at various stages of learning.

Let us consider κ_i = the dog and κ_j = DET NOUN. The pair (κ_i, κ_j) constitutes an abstraction chain as κ_j is a schematic instance of κ_i.

Using a distributional analysis, we can capture how the contexts of κ_i and κ_j vary, and how this variation is associated with **generalization**. If their cosine similarity decreases during training, it means that their contexts become more and more dissimilar: the model produces *DET NOUN* in new contexts which do not perfectly overlap with those of *the dog*, indicating that the network's babbling is becoming

1 The Jaccard index between CHILDES and OpenSubtitles remains above 0.5, even when considering the top 1M catenae, while the same index computed between CHILDES and Simple Wikipedia drops to around 0.13.

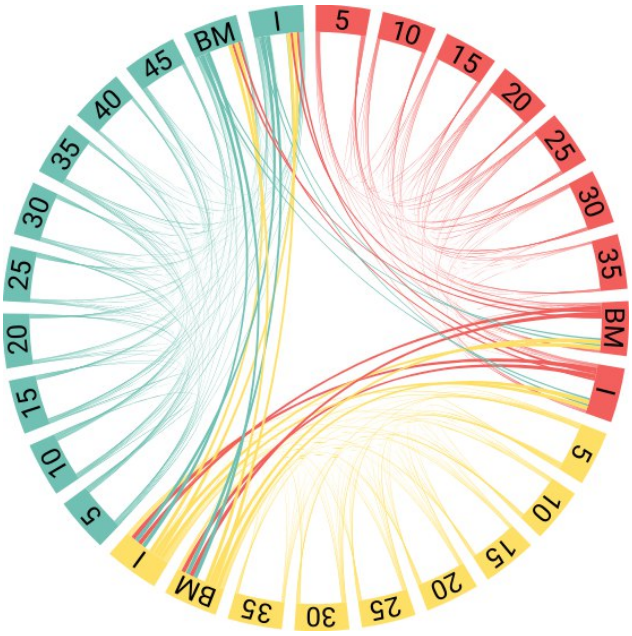


Figure 4.3: Correlation values (Spearman ρ) over top 10K catenae for each corpus (OpenSubtitles in green on the left of the plot, CHILDES in red in the top right, and Simple Wikipedia in yellow at the bottom) compared to the respective babbling (at intermediate stages of learning) and the best models (BM). The thickness of the connections is inversely proportional to correlation.

more productive (a graphical representation is given in Figure 4.4). In this case, we theorize that κ_j has been recognised as a partially independent pattern from κ_i. If, on the contrary, their cosine similarity increases, we might deduce that the network has recognized κ_j as partly unnecessary: it is correcting an overgeneralization.

For this analysis, vector spaces are created both for the input and for the *babbling* produced at each *acquisition step*. We consider catenae composed by 2 or 3 elements as both targets and contexts, and define co-occurrence as the presence of two catenae in the same sentence. Co-occurrences are weighted with PPMI and the space reduced to 300 dimensions with SVD (Klema and Laub, 1980).

In the first (and published) version of this study, the spaces were created using the DISSECT (Dinu et al., 2013) toolkit.

Some examples of pairs and their distributional similarities in each construction is shown in Table 4.12.

We then split catenae in three bins based on their *average distributional shift* and investigate the influence of input similarity over the schematization process of a construction.

Both for κ_i and κ_j, some *average distributional shifts* are shown in Table 4.13.

The hypothesis that we test is that constructions that underwent the highest shifts during training are those showing intermediate levels of similarities in the input distributional space (Section 4.1.4). To test it, we perform a *Kruskall-Wallis one-way analysis of variance test*, that turns out to be significant for groupings made on both κ_i and κ_j lists.

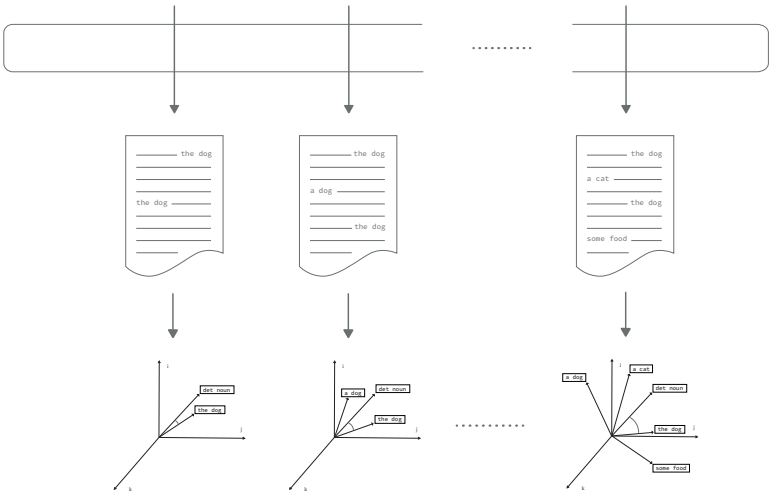


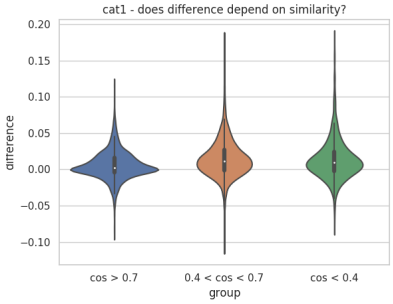
Figure 4.4: The figure shows our hypothesis: the *schematization* path for the construction *DET NOUN* can be described by the increased cosine distance between the vector representing the construction and its more prototypical instance (in child-directed data), which we represent as the vector *the dog*. The construction will emerge as a schematic pattern once more lexicalized instances will be produced and this variety is reflected in distributional vector space.

The test, in fact, returns a value of $p = 6.988142426844016\text{e-}28$ for κ_i and $p = 7.420868598608134\text{e-}32$ for κ_j. The result is confirmed by *Dunn's posthoc test*, as shown by Figure 4.5a, Figure 4.5b, Figure 4.5c, Figure 4.5d and Table 4.14, Table 4.17, Table 4.15 and Table 4.17.

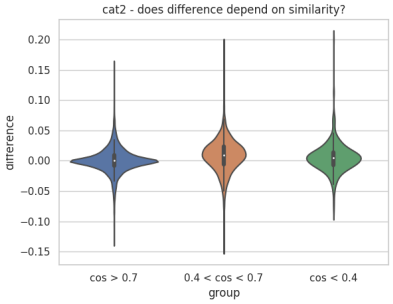
4.4 DISCUSSION

In line with the existing usage-based computational accounts, our experiments highlight how the CALaMo methodology can be deployed to evaluate the level of productivity of an LSTM trained on limited, child-directed data, using inspirations from constructionist approaches.

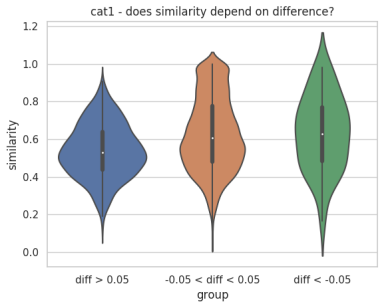
We have been able to show that Neural Language Models approximate the distribution of constructions at a quite refined level when trained over a bare 3M words from the CHILDES corpus, reproducing the distribution of grammatical patterns even when they are not fully lexicalized. The analysis in Section 4.2 indicates that the linguistic variety of OpenSubtitles is a potentially relevant benchmark to further investigate language acquisition, due to its similarity to the CHILDES data. In contrast, Simple Wikipedia has proved to be dissimilar to child-directed speech. This large difference should be taken into consideration when it comes to evaluating the grammatical abilities on the network: many of the studies cited in Chapter 2 use models trained on Wikipedia or similar varieties, which may complicate the acquisition



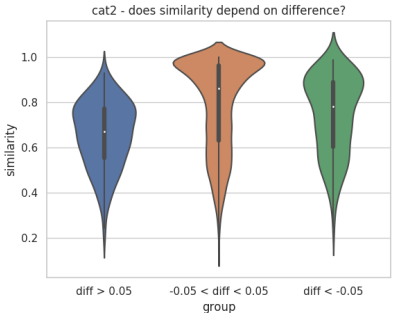
(a) Distribution of average distributional shifts for the three groups of κ_i for the three groups of constructions with low, intermediate and high average cosine similarities.



(b) Distribution of average distributional shifts for the three groups of κ_j for the three groups of constructions with low, intermediate and high average cosine similarities.



(c) Distribution of average cosine similarity for the three groups of κ_i for the three groups of constructions with low, intermediate and high average distributional shifts.



(d) Distribution of average cosine similarity for the three groups of κ_J for the three groups of constructions with low, intermediate and high average distributional shifts.

Figure 4.5

of generic grammatical phenomena heavily present in child-directed language. The analysis in Section 4.3 further illustrated how we can follow paths of abstraction by putting our grammar formalism in a vector space. Additional investigations are of course needed to confirm our results. In particular, we would like to target the behavior of some specific sets of structures.

Most importantly, the introduced methodology, despite being preliminary, presents a number of features that make our study fit in the usage-based theoretical framework while also using neural networks as language modeling tools, more specifically:

(I) it posits no sharp distinction between lexicon and grammar: fully lexicalized, partially filled and purely syntactic patterns are all part of our construction and can play a similar role in production. Different items can therefore be represented compared, irrespective of their *lexical* nature;

(II) it makes no assumption about the stability of the constructicon: what is relevant for productivity at the earliest stages of learning might become superfluous later on;

(III) all items in the constructicon are seen as form-meaning pairs (i.e., constructions by definition, as in Goldberg, 2006): a novel way of modeling constructional meaning is therefore introduced and represents a promising path for future studies;

(IV) distributional semantics is used both as a powerful quantitative tool and as a usage-based cognitive hypothesis, which leads us to specific assumptions about the cognitive format and origin of semantic representations (Lenci, 2008), and seems in line with the view of constructions as "invitations to form categories" (Goldberg, 2019).

Finally, we must account for potential biases introduced by applying dependency parsing to both input data and neural babbling: while this step is necessary to extract catenae, it introduces a non-negligible amount of noise, as the available pipelines are typically trained on different varieties than the ones considered in this study. In particular, the parser is somehow projecting its own categories, which have been acquired in a different setting and probably on a different variety, on our data. This currently limits the transferability of our results.

	CHILDES	opensubtitles	simplewiki
input	you tinker tot	this is no way to treat a lion !	She is the rector of the National Autonomous University of Honduras (UNAH) since 2009 .
	let 's see if I can make this turn here	re-entry into earth 's atmosphere in 37 minutes .	
	that 's Jim's business	are you worried i'm going to try and stop you?	
best model	she 's a fire	some of horases are here down .	She is a former municipality in the center of an arrondissement in the southwest of France .
	if I put it down for a snack	just lost it all .	
	oh I love you	i said we've ... lucky .	

Table 4.3: Examples from input text and babbling produced by the best model, for each corpus. Sentences have been sampled according to the distribution of sentence lengths in the data.

	CHILDES	OPENSUBTITLES	SIMPLEWIKIPEDIA
λ_1^X	on oking up in	by ground paper .	The 1872s match .
	he 's teeting tapion to close	uh , what is it	Things only regular main metal agreement with the problem .
λ_2^X	oh that 's silly	where do you think	Because he founded this unique community .
	I do n't know	that means he 'll see how they saw you .	Charles Operation
λ_3^X	ng have n't you	i want to see her .	He then breeds gold fishing and did not kill Asia so Top teases and both China by the time .
	look	wef golden statue rink , where you are , a pork .	There is other China shortesting .
λ_4^X	those things are coming and have webbles outside	it worked , and-we 're getting back on town .	Lales , who can breed together to be applied to White Dobest .
	I wan na take it off	that 's , dad .	The park won the Kansas Chairman Island in 1955 .
λ_5^X	what 's this	ah , that 's the way now i 'll finish with you today –	The last name Kanshan called his Stanley C. Rod to Sinma Luge Rio River is a Saxony movie .
	that 's your train	would you pass him one	In 1869 used for the project movement in President Daycho and Robert II to discover of the life .
λ_6^X	good girl	there 's no top going , okay	Bloodress called data in the route to poetry at the University of Marie Caldo , Perland Dihtriots rate .
	did n't you yes here we hafta do then	none of my monkey	He orders starting and many people lived in the rifle .
λ_7^X	oh I 've never seen any pigs for his special lot do	i 've been lookin ' all kindly soon	It is found in the region Picardie in the Aisne department in the north of France .
	yeah and there 's one in the wardrobe	actually , you 're almost three months .	Knowenstores can bronze getting too bad with the stizaur in a coastline above 200 and 480 for a different region of many different stories , and those in the Soviet Union .
λ_8^X		me too .	
		i could write what we need under the floor from there .	
λ_9^X		our army has created the theater at the slag of nightgown in government teen	
		why is there always one cone in my power and minawake-luther	

Table 4.4: Examples of babblings at different acquisition steps

	CHILDES								
	λ_1^C	λ_2^C	λ_3^C	λ_4^C	λ_5^C	λ_6^C	λ_7^C	BM	λ_C
λ_1^C	1	0,922	0,925	0,892	0,894	0,882	0,924	0,914	0,897
λ_2^C	0,919	1	0,93	0,935	0,924	0,917	0,946	0,951	0,924
λ_3^C	0,912	0,923	1	0,925	0,929	0,92	0,955	0,949	0,934
λ_4^C	0,887	0,929	0,916	1	0,894	0,885	0,94	0,942	0,921
λ_5^C	0,877	0,912	0,919	0,902	1	0,941	0,947	0,945	0,923
λ_6^C	0,87	0,912	0,92	0,899	0,938	1	0,94	0,946	0,936
λ_7^C	0,921	0,944	0,952	0,945	0,95	0,942	1	0,983	0,961
BM	0,906	0,945	0,945	0,946	0,949	0,946	0,983	1	0,96
λ_C	0,89	0,908	0,931	0,911	0,913	0,936	0,954	0,953	1

Table 4.5: Correlation values (Spearman ρ) computed over top $10K$ catenae extracted at each *acquisition step* for CHILDES corpus

	OPENSUBTITLES										
	λ_1^O	λ_2^O	λ_3^O	λ_4^O	λ_5^O	λ_6^O	λ_7^O	λ_8^O	λ_9^O	BM	λ_O
λ_1^O	1	0,851	0,883	0,907	0,879	0,892	0,911	0,913	0,907	0,915	0,87
λ_2^O	0,875	1	0,91	0,897	0,912	0,908	0,921	0,928	0,927	0,932	0,896
λ_3^O	0,871	0,894	1	0,904	0,918	0,929	0,933	0,938	0,911	0,94	0,894
λ_4^O	0,916	0,882	0,918	1	0,942	0,924	0,944	0,958	0,94	0,954	0,931
λ_5^O	0,903	0,906	0,934	0,951	1	0,938	0,951	0,966	0,954	0,961	0,942
λ_6^O	0,888	0,905	0,942	0,914	0,929	1	0,941	0,953	0,916	0,946	0,918
λ_7^O	0,881	0,832	0,906	0,896	0,873	0,906	1	0,935	0,946	0,948	0,884
λ_8^O	0,914	0,905	0,943	0,957	0,952	0,958	0,973	1	0,966	0,99	0,952
λ_9^O	0,905	0,829	0,911	0,906	0,874	0,908	0,978	0,953	1	0,978	0,891
BM	0,915	0,89	0,944	0,944	0,931	0,948	0,983	0,988	0,98	1	0,937
λ_O	0,877	0,892	0,895	0,936	0,94	0,924	0,927	0,956	0,928	0,949	1

Table 4.6: Correlation values (Spearman ρ) computed over top $10K$ catenae extracted at each *acquisition step* for Opensubtitles corpus

	SIMPLEWIKIPEDIA								
	λ_1^S	λ_2^S	λ_3^S	λ_4^S	λ_5^S	λ_6^S	λ_7^S	BM	λ_S
λ_1^S	1	0,882	0,897	0,893	0,875	0,879	0,894	0,887	0,843
λ_2^S	0,878	1	0,919	0,909	0,909	0,909	0,952	0,908	0,884
λ_3^S	0,899	0,918	1	0,908	0,916	0,927	0,954	0,956	0,912
λ_4^S	0,856	0,914	0,91	1	0,928	0,93	0,937	0,913	0,875
λ_5^S	0,822	0,905	0,908	0,927	1	0,967	0,925	0,928	0,871
λ_6^S	0,815	0,897	0,907	0,91	0,961	1	0,919	0,921	0,873
λ_7^S	0,896	0,948	0,955	0,934	0,929	0,937	1	0,946	0,923
BM	0,878	0,9	0,957	0,91	0,926	0,917	0,942	1	0,911
λ_S	0,856	0,872	0,901	0,868	0,87	0,874	0,914	0,900	1

Table 4.7: Correlation values (Spearman ρ) computed over top $10K$ catenae extracted at each *acquisition step* for SimpleWikipedia corpus

		CHILDES		OPENSUBTITLES		SIMPLEWIKIPEDIA	
		BM	λ_C	BM	λ_O	BM	λ_S
CHILDES	BM	1	0,96	0,632	0,62	0,283	0,276
	λ_C	0,953	1	0,646	0,636	0,285	0,275
OPENSUBTITLES	BM	0,712	0,708	1	0,937	0,315	0,311
	λ_O	0,729	0,739	0,949	1	0,347	0,344
SIMPLEWIKIPEDIA	BM	0,317	0,307	0,346	0,342	1	0,911
	λ_S	0,319	0,304	0,343	0,342	0,900	1

Table 4.8: Correlation values (Spearman ρ) computed over top $10K$ catenae extracted from each *input* and best performing model

catena	frequency	mi
largest MI		
@nsubj @root	294.59K	633.93K
@nsubj _VERB	269.81K	621.08K
_DET _NOUN	189.97K	552.32K
@det _NOUN	185.64K	550.92K
_VERB @obj	190.72K	520.82K
_PRON _VERB	271.44K	503.17K
_PRON @root	290.78K	487.42K
@nsubj _AUX @root	129.60K	478.86K
@nsubj _VERB @obj	111.34K	466.75K
_VERB _ADP @obl	68.30K	429.38K
MI near zero		
_NOUN @root @xcomp _NOUN	320	0.04
@cc _PRON @root you	169	0.03
@nsubj _ADV @det _NOUN _VERB	113	-0.01
_PRON want _ADJ	100	0.03
_ADV _AUX _VERB @advmod _NOUN	90	-0.05
_NOUN _NOUN @root _ADJ @obj	79	-0.01
_PRON _VERB @nsubj @aux _ADJ	78	0.03
_VERB _VERB _VERB _VERB @xcomp	77	0
@obl:tmod @aux _VERB	63	-0.03
@nsubj _PART _VERB _PART @root	53	-0.01
smallest MI		
_PRON _PRON	12.74K	-37.53K
_PRON @nsubj	17.50K	-35.54K
@root @nsubj	27.61K	-34.89K
@nsubj _PRON	11.63K	-30.47K
_PRON _AUX	19.00K	-27.03K
_VERB @nsubj	12.79K	-26.82K
_AUX _PRON	15.75K	-26.67K
_VERB @root	8.40K	-25.63K
_NOUN _PRON	8.01K	-24.28K
@root _AUX	21.86K	-24.27K

Table 4.9: Catenae extracted from CHILDES with their frequency and mutual information.

catena	frequency	mi
largest MI		
@nsubj _AUX @root	115.74K	466.29K
@nsubj _VERB	192.29K	454.95K
@nsubj @root	202.63K	425.80K
_VERB @obj	143.34K	405.82K
_DET _NOUN	146.24K	382.79K
@det _NOUN	142.79K	378.51K
_PRON _AUX @root	107.77K	367.34K
@nsubj @aux _VERB	77.18K	361.89K
_VERB _ADP @obl	60.26K	357.30K
_VERB @case @obl	58.17K	347.78K
MI around zero		
@root @obj _PRON @det _NOUN	200	0.05
i @root _ADP _ADV	156	0
_DET @nsubj _VERB _PRON _ADV	141	0.02
we here	96	0.05
_VERB @amod _NOUN _NOUN _ADV	77	0
@nsubj @root _PRON _PART _ADJ	75	0.04
@advcl @aux @nsubj @root	68	-0.03
_DET @obl you _VERB	62	-0.01
gets _NOUN	61	0
away @obl	59	0.01
smallest MI		
_PRON @nsubj	10.05K	-22.43K
@root @nsubj	9.93K	-22.34K
_PRON _PRON	6.01K	-20.97K
_NOUN _PRON	5.69K	-19.56K
_VERB @root	6.09K	-19.07K
@nsubj _PRON	5.71K	-17.39K
_NOUN @root	48.34K	-16.99K
_PRON _AUX	6.90K	-16.22K
_VERB @nsubj	6.13K	-15.97K
@root _AUX	6.44K	-15.79K

Table 4.10: Catenae extracted from opensubtitles with their frequency and mutual information.

catena	frequency	mi
largest MI		
_VERB _ADP @obl	76.03K	386.69K
_VERB @case @obl	75.77K	381.07K
@nsubj:pass @aux:pass _VERB _ADP @obl	24.27K	314.61K
@nsubj:pass @aux:pass _VERB @case @obl	24.20K	312.29K
_DET _NOUN	150.28K	311.67K
@det _NOUN	149.04K	309.86K
@nsubj:pass _AUX _VERB _ADP @obl	26.43K	302.78K
@nsubj:pass _AUX _VERB @case @obl	26.35K	300.35K
@nsubj @root	96.50K	299.59K
_DET _NOUN @case @nmod	51.00K	295.75K
MI around zero		
@det _PROPN _ADP _PROPN _NOUN	545	-0.04
_PROPN _PROPN @nsubj _NOUN _VERB	293	0.02
@det _NOUN _PROPN _AUX _PROPN	270	0.02
@det @obl @root @nmod	195	-0.03
_ADP _NUM _NOUN _ADP _PROPN	194	0.03
_PROPN is @compound _PROPN	178	0.02
@nsubj @det _NOUN @case @obl	145	0.03
@root _DET _NOUN @case _PRON	142	0.04
_AUX @root _ADP _ADP _PROPN	130	-0.04
_DET _PROPN @compound @nmod _NOUN	127	-0.01
smallest MI		
_PROPN _NOUN	34.51K	-13.55K
_PROPN @case	4.77K	-12.51K
_NOUN @nsubj	6.02K	-12.44K
_NOUN _ADJ	8.04K	-12.13K
_NOUN @obl	8.49K	-11.03K
_NOUN @case	2.32K	-9.78K
@nmod _NOUN	8.38K	-9.70K
@case @root	4.27K	-9.25K
_ADP @root	4.26K	-9.02K
_NOUN @compound	3.02K	-8.89K

Table 4.11: Catenae extracted from simplewikipedia with their frequency and mutual information.

κ_i	κ_j	λ_C	**BM**	λ_1^C	λ_2^C	λ_3^C	λ_4^C	λ_5^C	λ_6^C	λ_7^C	**distributional shift**
a minute	a _NOUN	0.28	0.32	0.71	0.51	0.44	0.39	0.38	0.37	0.34	0.37
a minute	a @root	0.13	0.19	0.49	0.37	0.26	0.20	0.21	0.22	0.20	0.30
you _VERB it	_PRON @root @expl	0.10	0.19	0.46	0.28	0.25	0.25	0.19	0.17	0.21	0.25
you _VERB you	you _VERB @iobj	0.28	0.40	0.68	0.56	0.47	0.49	0.39	0.42	0.43	0.25
we can _VERB	_PRON can @root	0.51	0.54	0.79	0.74	0.59	0.54	0.55	0.61	0.57	0.22
go _VERB @obj	_VERB @conj @obj	0.64	0.72	0.56	0.74	0.70	0.74	0.72	0.72	0.72	-0.16
_AUX hungry	@cop @conj	0.68	0.52	0.36	0.39	0.44	0.45	0.47	0.42	0.59	-0.24
can get	can @advcl	0.55	0.54	0.24	0.36	0.45	0.48	0.43	0.39	0.52	-0.28

Table 4.12: Pairs of catenae (κ_i, κ_j), their cosine similarity in the space obtained from CHILDES, in the space obtained from the best model (BM) and at all *acquisition steps*. The last column shows the difference between cosine similarity in λ_1 and cosine similarity in λ_7.

κ_i	**shift**	**cosine**	κ_j	**shift**	**cosine**
@nsubj @root so	0.18	0.43	more @root	0.2	0.21
@nsubj only @root	0.18	0.41	_AUX know @obj	0.19	0.66
what @root @obj	0.18	0.39	@advmod tell	0.17	0.64
what @advmod _VERB	0.16	0.19	@aux know @obj	0.16	0.71
only @root	0.16	0.38	@advmod can _VERB	0.15	0.76
more @root	0.16	0.23	know @obj	0.15	0.62
@root it @xcomp	0.15	0.61	a _NOUN	0.13	0.52
@det minute	0.15	0.25	might @root	0.13	0.70
_PRON only @root	0.15	0.53	_PRON @root n't	0.12	0.53
_VERB _DET minute	0.15	0.33	@root that _VERB	0.12	0.65
_PRON @root so	0.14	0.54	_VERB 'll @ccomp	0.12	0.71
_DET minute	0.134	0.33	_VERB me @obl	0.12	0.76

Table 4.13: Constructions with highest average shifts.

	negative	none	positive
negative	-	7.32e-01	1.33e-03
none	0.732	-	5.71e-29
positive	0.001	5.71e-29	-

Table 4.14: Posthoc significance tests results for Figure 4.5a

	negative	none	positive
negative	-	6.83e-06	4.57e-05
none	0.000	-	4.15e-29
positive	0.000	4.15e-29	-

Table 4.15: Posthoc significance tests results for Figure 4.5b

	$\cos < 0.4$	$0.4 < \cos < 0.7$	$\cos > 0.7$
$\cos < 0.4$	-	1.30e-02	1.53e-19
$0.4 < \cos < 0.7$	1.30e-02	-	5.07e-67
$\cos > 0.7$	1.53e-19	5.07e-67	-

Table 4.16: Posthoc significance tests results for Figure 4.5c

	$\cos < 0.4$	$0.4 < \cos < 0.7$	$\cos > 0.7$
$\cos < 0.4$	-	2.30e-04	1.83e-05
$0.4 < \cos < 0.7$	2.30e-04	-	4.158696e-73
$\cos > 0.7$	1.83e-05	4.16e-73	-

Table 4.17: Posthoc significance tests results for Figure 4.5d

5

POPULATIONS OF ARTIFICIAL SPEAKERS

In the nativist tradition, *linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly* (Chomsky, 1965).

The fact that speakers might not all exhibit the very same competence is now generally acknowledged, also in the nativist theory. In computational settings also, more attention is being given to variability. In computational linguistics, in particular, where datasets or annotations are routinely collected from sets of speakers, inter-annotator agreement (Artstein and Poesio, 2008) is considered a valuable parameter. But in spite of progress in that direction, during the evaluation of Neural Language Models, inter-speaker variability is hardly ever considered as a feature: models are never conceptualized as individual speakers, but rather as average representations of a homogeneous linguistic community. Their performances are therefore often compared to the average human performance, treating variability as noise rather than as an intrinsic component of the definition of language itself.

This bias probably stems from the opposing views on judgements of grammaticality that have emerged during linguistic history. The proponents of grammaticality as a graded property are to be found within the cognitive and usage-based tradition. Lakoff, 1973 states for instance:

- rules of grammar do not simply apply or fail to apply, they rather apply to a degree;
- grammatical phenomena form hierarchies which are largely constant from speaker to speaker, and in many cases, from language to language;
- different speakers (and different languages) will have different acceptability thresholds along these hierarchies.

The role of the linguistic community becomes more and more important as we exit the area of what is considered *core* grammatical knowledge: fully grammatical sentences (i.e., sentences that all speakers agree on being perfectly acceptable) can be judged without paying

very much attention to meaningfulness and interpretability (see the classic *colorless green ideas sleep furiously* example). On the other hand, interpretability, and cued meaning with it, becomes fundamental in judging cases on the continuum from full grammaticality to ungrammaticality (C. T. Schütze, 2015). So in order to fully explain actual speaker performance, as well as the speaker-listener interface, it is important to model linguistic diversity. Note that the point is not to define individual idiolects. But exploring the range of possibilities that are realized within a community is crucial to the definition of language and this is a highly underrepresented aspect in the approach to Neural Language Models' evaluation.

In this section, we use the CALaMo framework to make a new step in the evaluation of Neural Language Models. Our main contribution is to take seriously the idea of an individual speaker, and port it to a computational setting. That is, we will assume that every instance of a trained model can be regarded as a different individual in a population. While this idea can be found in other areas of the computational literature (for instance, in computational work on language evolution), we do not know of any prior study that would analyze the notion of grammar using artificial populations of speakers.

In the following, we will first show how to generate populations with different distributions of input. Then, on the basis of the generated speakers, we will address two theoretical points:

(I) we first set out to confirm the existence of a set of *core* constructions that are learned by all speakers of a community independently of the input they receive and in a statistically significant form;

(II) we then ask how speakers who have acquired diverse grammars can actually communicate: provided that a *core* set of constructions has been identified, we introduce a methodology to project it onto sentences and thus retrieve the shared signal in a speaker-listener interaction.

5.1 CALAMO FOR POPULATIONS: A FORMALIZATION

Using CALaMo, we introduce a methodology to compare the linguistic productions of a population of Neural Language Models. We train each model on a unique input, thus treating it as an individual speaker, and on this basis, explore the variation that emerges within the population.

In the rest of this chapter, we will use two different settings to generate our populations, resulting in:

- an exploration of a *homogeneous* population of 10 Neural Language Models, each trained on a 1-million subset of the CHILDES corpus;

- an exploration of an *inhomogeneous* population of 9 Neural Language Models, composed of 3 subgroups of 3 NLMs each, trained on 1-million words samples of CHILDES, OpenSubtitles and Simple Wikipedia respectively.

Having these two different settings will allow us to highlight how much diversity is created in a population when speakers' inputs follow significantly different distributions, compared to a case where all speakers are exposed to very similar data.

But before we proceed with our experiment, we will first formalize our population generation process.

5.1.1 *The population model*

As specified in Section 3.2, our population of speakers can be defined as

$$\Pi = (\sigma_1, \sigma_2, \cdots, \sigma_P) \tag{5.1}$$

Following the previous experiment, each speaker σ_k is a Neural Language Model, more specifically a vanilla character-based LSTM, this time trained on 1 million words.

As stated previously, we consider two different scenarios:

- a population of 10 homogeneous speakers, which we refer to as Π_H
- a population of 9 inhomogeneous speakers, which we refer to as Π_I

The grammatical conventions adopted within the communities are:

$$\Lambda_{\Pi_H} = (\lambda_{\sigma_1}, \lambda_{\sigma_2}, \cdots, \lambda_{\sigma_{10}}) \tag{5.2}$$

and

$$\Lambda_{\Pi_I} = (\lambda_{\zeta_1}, \lambda_{\zeta_2}, \cdots, \lambda_{\zeta_9}) \tag{5.3}$$

respectively.

With this setting (Figure 5.1), we try to identify the locus of variation among different speakers, under the assumption that a set of *core* constructions will emerge in the population as a set of shared constructions among individuals.

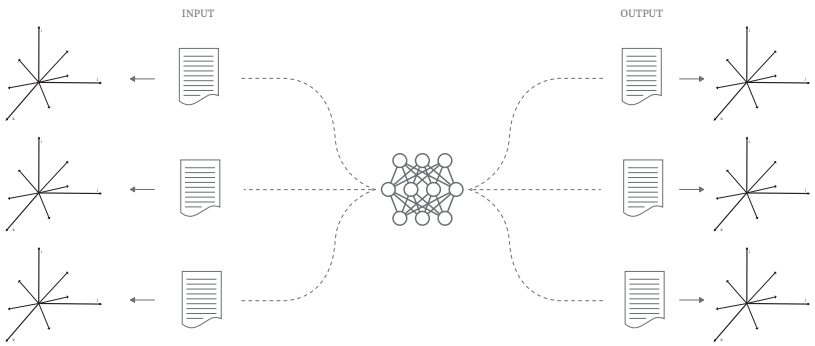


Figure 5.1: The figure shows how the CALaMo framework is adapted to a setting where a population of speakers is described. Each speaker receives a different input and, by means of the same acquisitional mechanism, produces an individual output.

5.1.2 *Data*

Each speaker σ_i in Π_H is trained on a randomly sampled extract ι_{σ_i} of the CHILDES corpus, each one containing 1 million words approximately.

As far as Π_I is concerned, instead, the population is composed of three subgroups, for which input data was selected as follows:

- ι_{ζ_1}, ι_{ζ_2}, ι_{ζ_3} correspond to ι_{σ_1}, ι_{σ_2}, ι_{σ_3} as $\zeta_{1\cdots3} \in \Pi_I$ correspond to speakers $\sigma_{1\cdots3} \in \Pi_H$;
- ι_{ζ_4}, ι_{ζ_5}, ι_{ζ_6} are randomly extracted samples of the Opensubtitles corpus, each one containing 1 million words approximately;
- ι_{ζ_7}, ι_{ζ_8}, ι_{ζ_9} are randomly extracted samples of the Simple Wikipedia corpus, each one containing 1 million words approximately.

Examples of input and 'babbling' output are reported in the next subsection. All text is then annotated by means of UDPipe (Section 3.3.3).

5.1.3 *Experimental pipeline*

For all speakers in the population, we build their constructions including the top 10K catenae retrieved in their babblings. As described in Chapter 3 and Chapter 4, catenae are weighted with an extended version of Mutual Information (Equation 4.2).

As far as **question (i)** is concerned, we perform the following steps. Let's consider the toy example provided in Figure 5.2.

Since different speakers are exposed to different input, different cases exist:

	speaker 1		speaker 2		speaker 3	
	input	babbling	input	babbling	input	babbling
cxn 1	43	1	70	1	38	2
cxn 2	93	2	93	2	57	3
cxn 3	89	3	56	3		
cxn 4	22	4			6	
cxn 5	30	5	16	7		
cxn 6	2	6	84	5	83	1
cxn 7	96	7	22	6	91	5
cxn 8	62	8		4	25	10
cxn 9	71	9	71		2	8
cxn 10	51	10	16			
cxn 11	76		8	8	89	
cxn 12			87	9	20	7
cxn 13			61	10	13	4
cxn 14	26		35		66	6
cxn 15			31		73	9

Figure 5.2: The figure shows a setting where, for each of the three considered speakers, their top associated 10 constructions are retrieved. This amounts for a total of 15 distinct constructions. For each of those, their rank is reported in columns labeled as *babbling* and their frequency in the input fed to each speaker in the columns labeled as *input*.

- a construction can be produced by all speakers, with all of them having been exposed to it in their input (e.g. *cxn 1* in the table);
- a construction can be produced by a subset of the speakers, with all speakers having been exposed to it through their input (e.g. *cxn 11*);
- a construction can be produced by a subset of the speakers, with some speaker not being exposed to it through their input (e.g. *cxn 3*).;

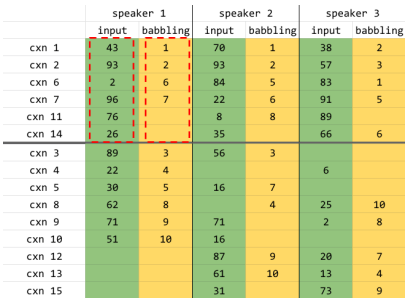
We restrict the analysis to those constructions that are present in the input of all speakers (see the top of Figure 5.3 above the line). The residual part of the table is however worth investigating, as we will show in Section 5.2.1, as it gives us a handle to quantify variations in the linguistic exposure of each individual.

We then check whether there exists a relation between the frequency of a catena in the input and the rank assigned through MI in the construction of the same speaker (Figure 5.4a) and of a different speaker (Figure 5.4b), under the assumption that high frequency will facilitate acquisition.

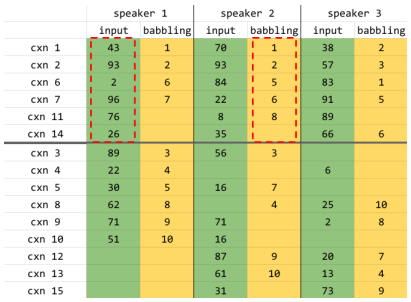
As it is evident from the example, not all speakers share the same constructions. We therefore build the set of core constructions as G_{10} for Π_H and G_9 for Π_I, and $G_{\leq 5}$for Π_H and $G_{\leq 4}$ for Π_I as the set of peripheral constructions.

	speaker 1		speaker 2		speaker 3	
	input	babbling	input	babbling	input	babbling
cxn 1	43	1	70	1	38	2
cxn 2	93	2	93	2	57	3
cxn 6	2	6	84	5	83	1
cxn 7	96	7	22	6	91	5
cxn 11	76		8	8	89	
cxn 14	26		35		66	6
cxn 3	89	3	56	3		
cxn 4	22	4			6	
cxn 5	30	5	16	7		
cxn 8	62	8		4	25	10
cxn 9	71	9	71		2	8
cxn 10	51	10	16			
cxn 12			87	9	20	7
cxn 13			61	10	13	4
cxn 15			31		73	9

Figure 5.3: Constructions from Figure 5.2 are now separated into the ones present in all speaker inputs (at the top), and others.



(a) Comparing the input frequency of shared constructions with the rank assigned by MI in a speaker's babbling.



(b) Comparing the input frequency of shared constructions with the rank assigned by MI across two speakers.

Figure 5.4

As detailed in Chapter 3:

$$G_{\geq p} = \left\{ \kappa \mid \sum_{i=0}^{P} X(\kappa, \sigma_i) \geq p \right\} \tag{5.4}$$

with

$$X(\kappa_i, \sigma_j) = \begin{cases} 1 & \text{if } \kappa \in \Lambda^{\sigma_j} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

being an auxiliary function that evaluates to 1 if the construction κ appears in the production of speaker σ_j and 0 otherwise (this just helps us count how many speakers use construction κ in their babbling).

Figure 5.5 shows how the groups would be formed in our toy example.

	speaker 1		speaker 2		speaker 3		
	input	babbling	input	babbling	input	babbling	
cxn 1	43	1	70	1	38	2	core
cxn 2	93	2	93	2	57	3	
cxn 6	2	6	84	5	83	1	
cxn 7	96	7	22	6	91	5	
cxn 11	76		8	8	89		periphery
cxn 14	26		35		66	6	

Figure 5.5: Core (top) and periphery (bottom) constructions, as extracted from the constructions in Figure 5.2.

Given the different sets, we check whether there is a significant difference in the input frequency for the three groups of constructions (the *core* group, the *periphery* group and the *residual*, i.e. the group of constructions that are neither core nor periphery).

Some cases appear to be particularly interesting: namely, core constructions that appear with low frequency in all or a subgroup of the population, and peripheral constructions that appear with high frequency (Figure 5.6 – in the example, frequency values are between 1 and 100). We qualitatively investigate these groups.

	speaker 1		speaker 2		speaker 3		
	input	babbling	input	babbling	input	babbling	
cxn 1	43	1	70	1	38	2	core
cxn 2	93	2	93	2	57	3	
cxn 6	2	6	84	5	83	1	
cxn 7	96	7	22	6	91	5	
cxn 11	76		8	8	89		periphery
cxn 14	26		35		66	6	

Figure 5.6: Outliers from the core and periphery groups of constructions. Speaker 1 displays a low frequency, core construction, while speaker 3 shows an example of a high frequency, periphery construction.

For **question (ii)**, we explore $\widetilde{\Lambda_G}$ for various definitions of G: in the case of Π_H, $\widetilde{\Lambda_{G_{10}}}$ represents the approximation of Λ obtained by considering only *core* constructions shared by the population. We introduce a function P_Λ that, given a sentence τ, produces τ_Λ being the representation of the sentence obtained by employing only the categories (i.e., the constructions) in Λ. The idea is that, by means of $\widetilde{\Lambda_G}$, we can take different perspective on the same sentence: we can see for instance what can be considered common ground by using the core constructions as G, or what is idiosyncratic of a specific speaker by using the constructions that he does not share with the whole population.

la | la | @flat:foreign | @flat:foreign
la | la | @flat:foreign | la
la | la | la | @flat:foreign
la | la | la | la
la | _X | @flat:foreign | @flat:foreign

Table 5.1: Catenae that are not shared by all speakers in Π_H. Only five are reported there but the other twenty show very similar patterns.

5.2 CORE AND PERIPHERY

In this section, we will detail the results obtained with respect to **question (i)**. Section 5.2.1 deals with the analysis of the top $10K$ constructions considered for each speaker in the population. Section 5.2.2 explores the relation between frequency in the input and the rank assigned to a construction by each speaker. Section 5.2.3 and Section 5.2.4 take a closer look at *core* and *periphery* constructions.

5.2.1 *Construction overlap across speakers*

As explained in Section 5.1.2, the two populations are composed by speakers trained on different language samples.

Some examples of the top ranked catenae in the construction of each speaker can be found in table below:

Proceeding with the analysis, we consider for each speaker their top $10k$ constructions, resulting in 11078 unique constructions across speakers for Π_H and 20124 unique catenae for Π_I. We then restrict ourselves to the subset of constructions to which all 10 speakers have been exposed through their input: this results in 11051 constructions out of 11078 for Π_H and 13732 out of 20124 for Π_I. As might be expected given the differences in sampled corpora, there is much more overlap among speakers in Π_H than in Π_I (close to 100% in the former case, and only 68% in the latter). The greater the variation in the input, the greater the difference among the constructions of the speakers will be.

In both cases, we observe constructions in the babbling which are not uniformly present in the input data fed to all learners. In the case of Π_H, they are reported in Table 5.1 and seem to be extracted from ill-formed sentences.

The case of Π_I is of course much more interesting: the catenae emerging from the babblings but missing from the input data for some speakers can be taken as an interesting proxy for the type of variations existing within the population.

We first notice that those 6392 catenae are uniformly distributed among the speakers, as reported in Table 5.2. This is interesting, because it shows that inter-speaker variability is quite *balanced*, meaning

ζ_1	3082	ζ_6	3115
ζ_2	3161	ζ_7	2715
ζ_3	3103	ζ_8	2727
ζ_4	3160	ζ_9	2679
ζ_5	3128		

Table 5.2: Number of catenae absent from each speaker's input data.

$\zeta_{1,2,3}$	childes	2800
$\zeta_{4,5,6}$	OpenSubtitles	2839
$\zeta_{7,8,9}$	Simple Wikipedia	2450

Table 5.3: Number of catenae absent from each group's input data.

that each speaker has on average been exposed to half of the catenae that some other speaker has missed in their training.

We then consider the constructions that exist in some of the babbling outputs but are absent from all input data. These are the same as those reported in Table 5.1 and they are likely to also come from ill-formed sentences.

We also notice (Table 5.3) that speakers trained on data sampled from the same distribution tend to miss, from their input, the same constructions. For instance, speakers $\zeta_{1,2,3}$, trained on CHILDES, miss 2800 catenae from their input, which accounts for approximately 90% of the number of constructions that each of them hasn't seen in their input individually. Examples of absent catenae for each group are reported in Table 5.4.

Finally, when looking at intersections among groups of speakers, we notice how speakers trained on CHILDES and speakers trained on OpenSubtitles miss from their input most of the overall discarded catenae, while speakers from Simple Wikipedia show a different pattern. Out of the 6392 discarded constructions, in fact, 2647 are absent from the input fed to speakers $\zeta_{1,2,3,4,5,6}$, meaning that these constructions are likely to be found in input data sampled from Simple Wikipedia.

5.2.2 *Relation between input frequency and babbling rank*

Having quantified the overlap between speakers and domains, we check whether the frequency of a catena in the input is related to the rank assigned through MI in the construction built from babbling data of the same speaker. The assumption is that core constructions should be frequent in the input data, and therefore lend themselves to acquisition regardless of the specific distribution presented to the speaker.

Childes	Simple Wikipedia	OpenSubtitles
@case | United | _PROPN	_NOUN | of | @nmod | _PROPN	@nsubj | _VERB | you | @ccomp
@nsubj:pass | @aux:pass | called | _PROPN	@case | United | _PROPN	and | @nsubj | @cop | @root
_DET | United | States	of | _DET | _PROPN | _PROPN	_PRON | gon | _VERB
of | _DET | _PROPN | _PROPN	_DET | United | States	we | _AUX | _PART | _VERB
@case | @det | States	@nsubj:pass | @aux:pass | called | _PROPN	you | _AUX | @root | @obj
In | @obl | @root	@case | @det | States	i | 'm | _NOUN
_NOUN | of | _PROPN | _PROPN	In | @obl | @root	do | @nsubj | @root | @obj
_PRON | @aux:pass | born | @obl	_NOUN | of | _PROPN | _PROPN	@advmod | you | _VERB | @obj
@nsubj:pass | found | @compound | _NOUN	_PRON | @aux:pass | born | @obl	@root | na | _VERB | @obj
@nsubj:pass | @root | _PROPN | _NOUN	@nsubj:pass | found | @compound | _NOUN	i | @root | @det | _NOUN
The | @root	@nsubj:pass | @root | _PROPN | _NOUN	@nsubj | 'll | @root | _VERB
@nsubj | a | commune	The | @root	i | @root | _ADP | _NOUN
@obl | _ADP | _PROPN | _PROPN	@nsubj | a | commune	@aux | _PRON | _VERB | @xcomp
_AUX | @root | in | department	@obl | _ADP | _PROPN | _PROPN	@aux | @nsubj | _VERB | _PRON
_NOUN | _ADP | @nmod | France	_AUX | @root | in | department	@obj | are | you | doing
department | @det | _NOUN | France	_NOUN | _ADP | @nmod | France	@advmod | _X | @flat:foreign | @flat:foreign
found | @case | @compound | @obl	department | @det | _NOUN | France	do | n't | _PRON | _VERB
@advmod | _X | @flat:foreign | @flat:foreign	found | @case | @compound | @obl	i | @root | _NOUN | _NOUN
_PROPN | _PROPN | @compound | _PROPN	_PROPN | _VERB | the | @obj	n't | @root | @obj | _VERB
was | _VERB | @case | _PROPN	@advmod | _X | @flat:foreign | @flat:foreign	_PRON | think | _AUX | @ccomp

Table 5.4: Some constructions absent from the input of groups of speakers trained on the three domains (Childes, Simple Wikipedia, OpenSubtitles).

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}
σ_1	0.75	0.76	0.74	0.75	0.75	0.75	0.75	0.75	0.74	0.75
σ_2		0.75	0.74	0.75	0.75	0.75	0.75	0.74	0.74	0.75
σ_3			0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
σ_4				0.75	0.75	0.75	0.75	0.75	0.74	0.75
σ_5					0.75	0.75	0.75	0.75	0.74	0.75
σ_6						0.75	0.75	0.75	0.74	0.75
σ_7							0.75	0.74	0.74	0.75
σ_8								0.75	0.75	0.75
σ_9									0.74	0.75
σ_{10}										0.75

Table 5.5: Spearman correlation coefficients between input frequency (on the rows) and ranking assigned by each speaker in Π_H (on the columns). All correlations are negative (the higher the frequency, the lower the rank value, hence the higher the position in the ranked list). We omitted the minus for readability purposes.

Results, reported in Table 5.5 and Table 5.6, show that correlations (Spearman's ρ) are significant: the higher the frequency, the higher the position in the ranked list of catenae, both for Π_H and Π_I. However, while being generally high and significant (with the due exceptions for Π_I), the correlation figures are also not perfect, suggesting that input frequency is not the only relevant factor for a catena to be acquired and reproduced in the babbling.

5.2.3 *Core and periphery*

Following the notation introduced in Section 5.2.2, we then create, for Π_H:

- the set of *core* constructions $G_{10} = \left\{ \kappa \mid \sum_{i=0}^{1} 0X(\kappa, \sigma_i) = 10 \right\}$
- the set of constructions in the *periphery* as $G_{\leq 5} = \left\{ \kappa \mid \sum_{i=0}^{1} 0X(\kappa, \sigma_i) \leq 5 \right\}$

and similarly for Π_I:

- the set of *core* constructions $G_9 = \left\{ \kappa \mid \sum_{i=0}^{9} X(\kappa, \zeta_i) = 9 \right\}$
- the set of constructions in the *periphery* as $G_{\leq 4} = \left\{ \kappa \mid \sum_{i=0}^{9} X(\kappa, \zeta_i) \leq 4 \right\}$

As far as Π_H is concerned, being trained on random samples taken from the same distribution, the speakers share most of the constructions (9086 out of 11051). Periphery constructions are instead 1287 out of 11051 and the residual class contains 678 constructions.

	ζ_1	ζ_2	ζ_3	ζ_4	ζ_5	ζ_6	ζ_7	ζ_8	ζ_9
ζ_1	0.77	0.77	0.77	0.77	0.75	0.75	0.39	0.44	0.41
ζ_2		0.77	0.77	0.77	0.75	0.75	0.39	0.44	0.41
ζ_3			0.77	0.77	0.75	0.75	0.39	0.43	0.41
ζ_4				0.77	0.75	0.75	0.39	0.43	0.41
ζ_5					0.75	0.75	0.39	0.44	0.42
ζ_6						0.75	0.39	0.44	0.41
ζ_7							0.81	0.72	0.73
ζ_8								0.72	0.73
ζ_9									0.73

Table 5.6: Spearman correlation coefficients between input frequency (on the rows) and ranking assigned by each speaker in Π_I (on the columns). All correlations are negative (the higher the frequency, the lower the rank value, hence the higher the position in the ranked list). We omitted the minus for readability purposes.

The situation is different for Π_I, where we find 3277 *core* catenae and 6720 *peripheral* ones.

5.2.4 *Low-frequency core and high-frequency periphery constructions*

Next, we examined the distribution of *core* and *periphery* catenae with respect to input frequency for each speaker σ_i and ζ_j.

For Π_H, the boxplots in Figure 5.7 show significant differences among the three groups for all speakers. For Π_I, the same can be noticed (Figure 5.8).

In both cases, and especially in the case of Π_H, there appear to be two groups of outliers:

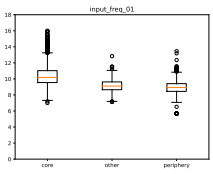
LFC low frequency *core* constructions

HFP high frequency *peripheral* constructions

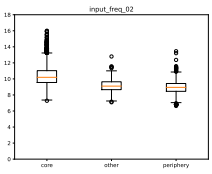
These appear to be particularly interesting as their examination might show why some catenae get picked up by all speakers despite their low frequency in the former case, and why some speakers, despite being exposed to a catena with a significant input frequency, do not pick it up.

We more formally define these outliers by computing, for each speaker, the first and third quartile and the interquartile range (reported in appendix) on the input frequencies: the constructions that fall outside of the respective quartile $+/-1.5$ times the interquartile range are considered outliers for that group.

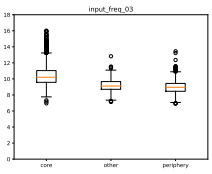
In formula (we consider the case of Π_H):



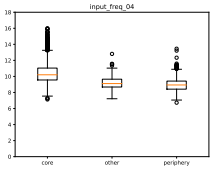
(a) Speaker σ_1



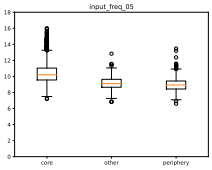
(b) Speaker σ_2



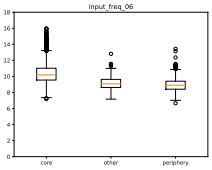
(c) Speaker σ_3



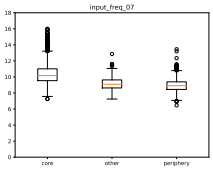
(d) Speaker σ_4



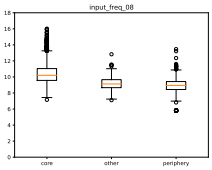
(e) Speaker σ_5



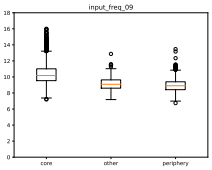
(f) Speaker σ_6



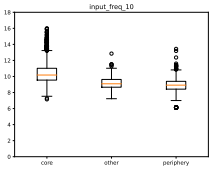
(g) Speaker σ_7



(h) Speaker σ_8

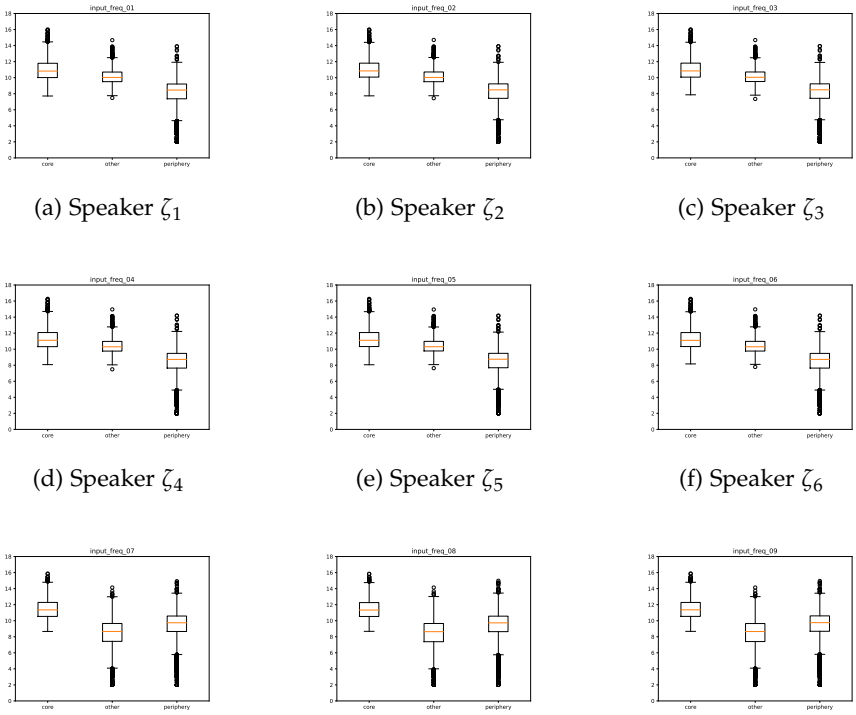


(i) Speaker σ_9



(j) Speaker σ_{10}

Figure 5.7: Differences in the input frequencies of constructions in the three groups (*core* on the left of each subfigure, *periphery* on the right of each subfigure and the residual group in the middle), for each speaker in Π_H



(a) Speaker ζ_1 (b) Speaker ζ_2 (c) Speaker ζ_3

(d) Speaker ζ_4 (e) Speaker ζ_5 (f) Speaker ζ_6

(g) Speaker ζ_7 (h) Speaker ζ_8 (i) Speaker ζ_9

Figure 5.8: Differences in the input frequencies of constructions in the three groups (*core* on the left of each subfigure, *periphery* on the right of each subfigure and the residual group in the middle), for each speaker in Π_I

9	gon na be @xcomp
9	a lot of @nmod
2	's gon na @xcomp

Table 5.7: Low frequency core constructions in Π_H. The table shows, on the left column, the number of speakers for which that construction qualifies as *low-frequency*

$$LFC_{\sigma_i} = \{\kappa \| \kappa \in G_{10} \wedge f(\kappa, I_{\sigma_i}) \leq q_1 - 1.5 * iqr\} \tag{5.6}$$

and

$$HFP_{\sigma_i} = \{\kappa \| \kappa \in G_{\leq 5} \wedge f(\kappa, I_{\sigma_i}) \geq q_3 + 1.5 * iqr\} \tag{5.7}$$

where *LFC* stands for *low frequency core* and *HFP* for *high frequency periphery* respectively, $f(\cdot, \cdot)$ is a function that returns the frequency of a catena in a given body of text, q_i indicates the quartile, and *iqr* stands for *interquartile range*.

By means of LFC_{σ_i} and HFP_{σ_i} we obtain, for each speaker σ_i, a set of the constructions that, while being in its construction, appeared in its input with *surprising* frequency values.

Looking at Π_H, we find some consistency among speakers with respect to which constructions are identified as outliers. For the *LFC* constructions, as shown in Table 5.7, 9 out of 10 speakers show the same two constructions as outliers. Similarly, for *HFP* constructions, most of the outliers are shared by the entire population of ten speakers (Table 5.8). Examples of sentences containing *LFC* constructions in the CHILDES corpus are shown in Figure 5.9 and Figure 5.10.

The case of Π_I is instead different. None of the speakers show any *LFC* constructions. In addition, as the figures show, the speakers have different sets of high-frequency core constructions. On the other hand, *HFP* constructions are shared by the majority of speakers (Table 5.9).

5.3 PERSPECTIVE AMONG SPEAKERS

We then turn to **question (ii)**, namely the issue of speaker perspective in the course of communication. We explore the input through *core* and *periphery* constructions, that is:

- $\widetilde{\Lambda_{G_{10}}}$ and $\widetilde{\Lambda_{G_{\leq 5}}}$ for P_H
- $\widetilde{\Lambda_{G_9}}$ and $\widetilde{\Lambda_{G_{\leq 4}}}$ for P_I

10	NOUN VERB NOUN NOUN
10	NOUN PRON AUX VERB
10	@root NOUN PRON VERB
10	VERB @advmod ADV
10	have NOUN
10	DET @root NOUN
10	VERB NOUN @obj
10	NOUN @nsubj VERB NOUN
10	@nsubj @case NOUN
10	PRON @root you
10	AUX NOUN
10	PRON the NOUN
10	PRON ADJ
10	VERB @obl VERB
10	VERB you
10	@root @obj @advmod
9	PRON AUX PRON VERB
5	AUX @root you
1	@det @nsubj VERB NOUN
1	@nsubj VERB PRON ADV
1	NOUN @cop ADJ

Table 5.8: High frequency periphery constructions in Π_H. The table shows, on the left column, the number of speakers for which that construction qualifies as *high-frequency*

ζ_1	ζ_2	ζ_3	ζ_4	ζ_5	ζ_6	ζ_7	ζ_8	ζ_9	
•	•	•	•	•	•	•	•	•	@nsubj NOUN
•	•	•	•	•	•	•	•	•	DET @nsubj
•	•	•	•	•	•	•	•	•	NOUN @root NOUN
•	•	•	•	•	•	•	•	•	NOUN @root
•	•	•	•	•	•	•	•	•	NOUN VERB
•	•	•	•	•	•	•	•	•	@det @nsubj
•	•	•	•	•	•				@root @advmod
•	•	•	•	•	•				NOUN PRON VERB
•	•	•	•	•	•				@root VERB PRON
•	•	•	•	•	•				PRON NOUN NOUN
						•	•	•	@root @case PROPN
						•	•	•	PROPN PROPN
						•	•	•	@case PROPN
						•	•	•	ADP PROPN
						•	•	•	@root ADP PROPN
						•	•	•	VERB PROPN
						•	•	•	@compound PROPN
						•	•	•	@root PROPN
						•	•	•	VERB PROPN PROPN
						•	•	•	ADP PROPN PROPN
						•	•	•	@case PROPN PROPN
						•	•	•	PROPN @root
	•			•					ADV VERB PRON
				•					the @root

Table 5.9: High frequency periphery constructions in Π_I. The table shows, on the left column, the number of speakers for which that construction qualifies as *high-frequency*

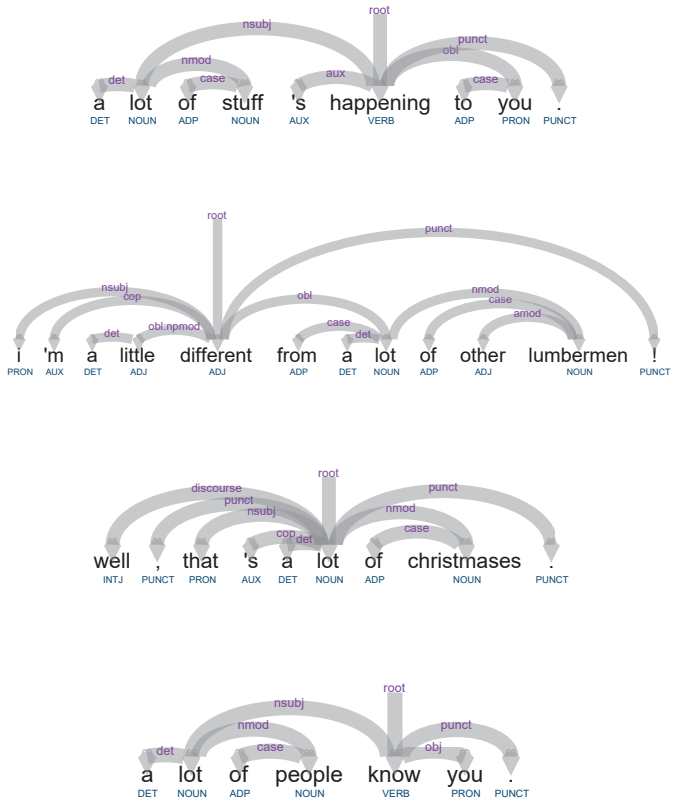


Figure 5.9: Sentences in the CHILDES corpus containing the construction *a lot of @nmod*

The fact that constructions are not entirely shared by speakers has consequences for the way we formalise communication. When a listener parses an utterance, we cannot assume that they retrieve the set of constructions that the speaker used to build their sentence. Most likely, they activate some shared and some idiosyncratic catenae during the listening process. The linguist who wishes to retrieve the part of the signal that was *actually* passed on needs a methodology to project core constructions onto sentences.

We hypothesize that *core* and *periphery* contribute differently to the meaning of a sentence: being shared at the population level, the *core* construction might help identifying, for instance, the type of event described by the sentence, or telling apart questions from declarative sentences, or any other stored linguistic knowledge that might be part of a community's common ground. *Peripheral* constructions, instead, as they are shared only by a subset of the population and are therefore in a way idiosyncratic to each speaker, help clarifying the commu-

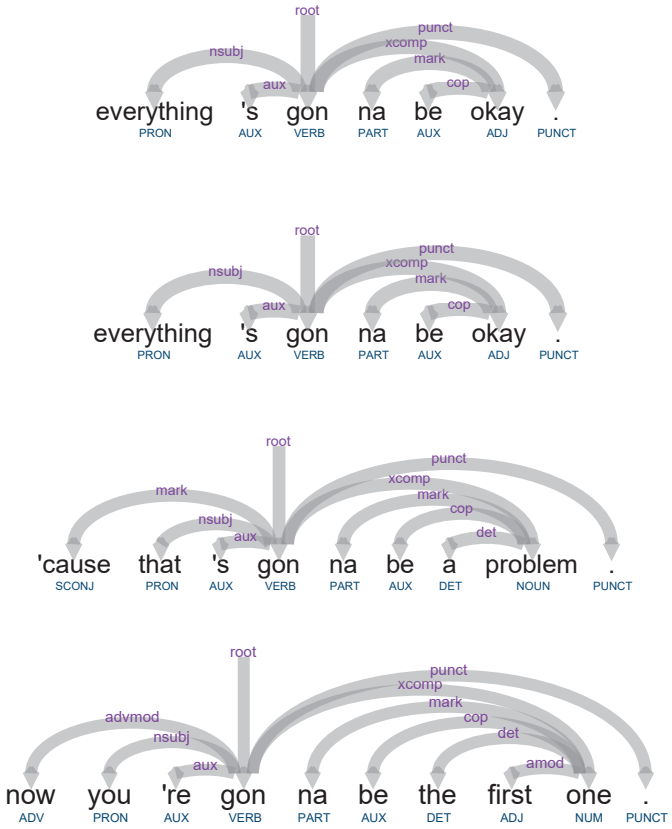


Figure 5.10: Sentences in the CHILDES corpus containing the construction *'s gon na @xcomp*

nicative intention of the speaker. In other words, a listener uses their most discriminative constructions to distinguish between the diverse meanings that different speakers may want to convey.

Computationally, this can be implemented as follows: given a set of sentences S, it is possible to translate them using either core constructions or peripheral constructions, thus obtaining two new sets of 'representations' for the set, C_S and P_S. We hypothesise that $|C_S| \ll |P_S|$ because a significant number of sentences will be projected onto the same core representation.

5.3.1 *The translation process*

Given a construction λ and a sentence s, we can retrieve from λ the set of constructions that get *activated* when building s. Constructions in λ are stored as tokens without any reference to their original positions in text. Therefore, it is possible that a single construction, such as

DET NOUN, gets activated multiple times in a sentence (e.g., *the cat is on the mat*). Similarly, the same construction can be activated by different sentences that however show very different shapes of that same construction: for instance, the *nsbj root* construction would look very different in the two sentences *Anne loves Barbara* and *the girl with long blond hair plays drums*.

What we want to achieve is a representation of the sentence (i.e., a *translation*) involving the activated constructions. For instance, if we activated the two constructions *nsbj root* and *root dobj* from a construction, we would obtain:

- Anne loves Barbara $\longrightarrow$ nsbj root dobj
- the girl with long blond hair plays drums $\longrightarrow$ X nsbj X X X X root dobj

Once the set of activated constructions in λ is identified, therefore, the first step we need to perform is to map each construction to its possible positions in the sentence s. This operation will produce a set of *sentominos* (a word modeled after the word *tetriminos* or *polyominos*. A high-level representation of the process is shown in Figure 5.11).

More formally, a *sentomino* can be defined as:

$$p =< p_1, \cdots, p_n > \tag{5.8}$$

where n is the length of the sentence s in terms of tokens, and

$$p_i = \varnothing \text{ OR } p_i = \kappa[j] \tag{5.9}$$

for a given construction κ.

For instance, for the sentence *Anne loves Barbara* and the construction *nsbj root*, the corresponding *sentomino* would be:

$$p =< \text{nsbj}, \text{root}, \varnothing > \tag{5.10}$$

Not all produced *sentominos* are compatible with one another, of course. Let us consider a very simple example for the sentence *Anne loves Barbara*: in this case, we can imagine the set of activated constructions as the set (*noun*, *nsbj*, *nsbj root*, *root noun*). Consequently, the set of *sentominos* would be composed by *noun* $\varnothing$ $\varnothing$, $\varnothing$ $\varnothing$ *noun*, *nsbj* $\varnothing$ $\varnothing$, *nsbj root* $\varnothing$, $\varnothing$ *root noun*.

We cannot, for instance, match the two *sentominos noun* $\varnothing$ $\varnothing$ and *nsbj* $\varnothing$ $\varnothing$ as, for the first token in the sentence, they contain different representations. On the other hand, the two *sentominos nsbj* $\varnothing$ $\varnothing$ and - *root noun* can be composed, resulting in *nsbj root noun*, which can be considered a *sentomino* itself and also a representation for the sentence (as no further operations are possible, all slots are filled).

Our aim is to build a representation of a sentence given the available *sentominos*. In order to do so, we need three further steps:

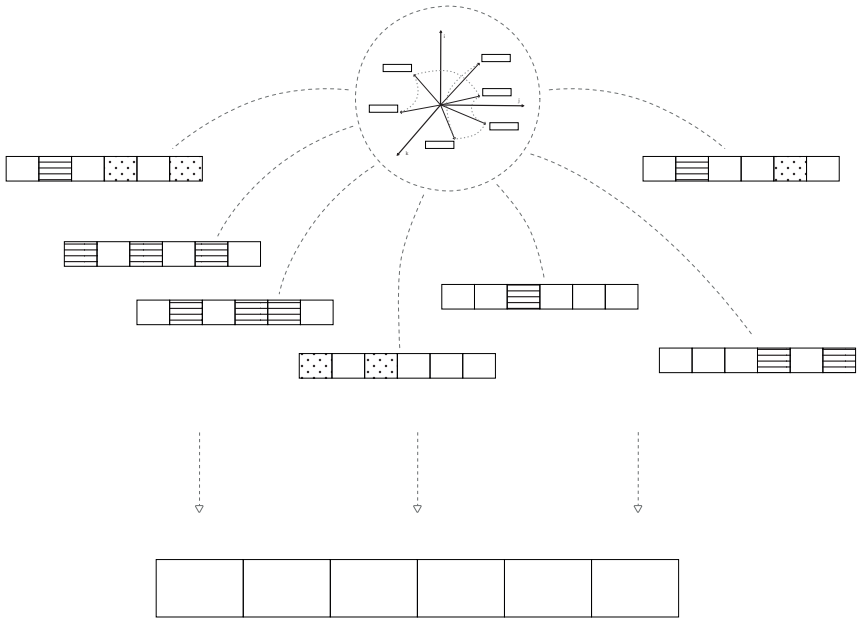


Figure 5.11: The picture illustrates how the translation process takes place. Given a sentence of length n (in this case, $n = 6$, the sentence is represented by the empty array at the bottom of the figure), and a construction (depicted at the top of the figure), a set of constructions get activated. These need to be *projected* on the full length of the sentence, as every subpart of the construction corresponds to a specific token in the sentence. These *projected* constructions are what we call *sentominos* and are represented in the middle of the picture. The different patterns represent the idea that constructions contain different kinds of lexical or categorical information.

(I) define a match function that determines whether two *sentominos* are compatible

(II) build all possible representations for the given sentence

(III) assign a score to each representation in order to identify the *best* one

5.3.1.1 *Define a match*

Let us define two *sentominos* for a sentence of size n:

$$p =< p_1, \cdots, p_n > \tag{5.11}$$

and

$$q =< q_1, \cdots, q_n > \tag{5.12}$$

In order to build a translation for s, we need to find out whether p and q can both contribute to the translation. In other words, if they are *matchable* or compatible with one another (Figure 5.12).

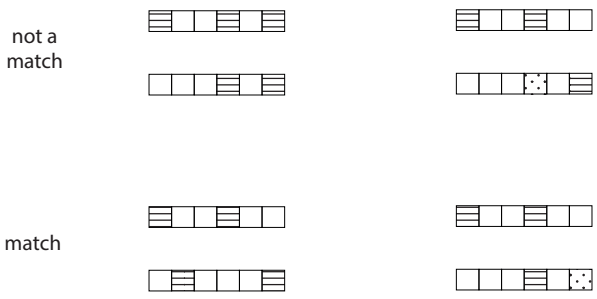


Figure 5.12: *Sentominos* are defined to match if they have no overlapping valorized position or if they contain matching values in all their common valorized positions (bottom part of the picture). On the other hand, they do not match if one is a subset of the other or if their common valorized positions do not match.

In order to define the matching function, we note that not all items in the *sentominos* will have a value. For instance, if the construction is composed by three elements, as for instance the *det adj noun* construction, when building a *sentomino* for a sentence of length n, only 3 out of n slots in the *sentomino* will contain a value, namely the slots corresponding to positions in the sentence where the determiner, the adjective and the noun are placed. We therefore consider I_p and I_q as the set of positions in p and q that actually contain some information.

Different cases can arise:

- if $I_p \cap I_q = \varnothing$, then p and q are matchable, as their values pertain to different positions in s

- if $I_p \cap I_q \neq \varnothing$, then we have to make sure that those values in the intersections are the same: $\forall i \in I_p \cap I_q \| p_i = q_i$
- as a further step, we do not want to consider p and q to be matchable if either $p \subset q$ or $q \subset p$

Obviously, the matching relation is symmetric, so if p is matchable with q, then q is matchable with p.

5.3.1.2 *Build a representation*

A full representation (or translation) s_λ for a sentence s can be defined as a set of *sentominos* $p^1, \cdots, p^k$.

By computing the matching relation between all possible *sentominos*, we define a graph of relation on the set of *sentominos* that can help us identify the subsets that can be collapsed in a sentence representation s_λ (Figure 5.13).

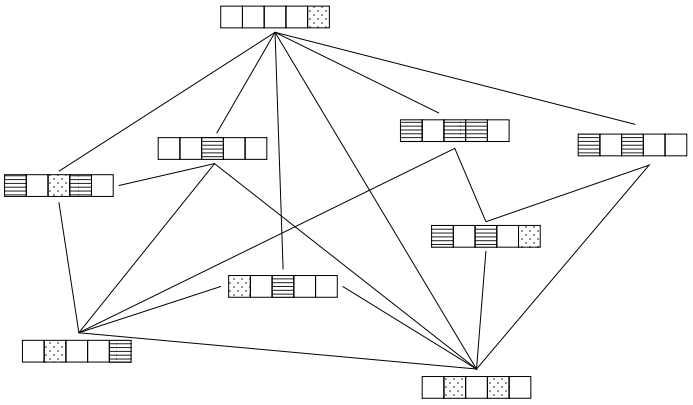


Figure 5.13: By means of the matching functions, we defined a graph containing all the activated *sentominos* for a sentence s. Edges on the graph indicate that two *sentominos* are compatible with one another.

We note that all the *sentominos* in s_λ have to be mutually matchable: what we are looking for are therefore cliques on the graph of relations. We actually restrict to maximal cliques, as we are looking for the representations that best (as in, most extensively, provide best coverage) approximate the sentence.

Formally, given a (undirected) graph, a *clique* can be defined as a subset of vertices such that every two vertices in the subset are adjacent (i.e., connected). A *maximal clique* is a clique that cannot be extended by adding any other vertex in the graph. An intuitive representation is given in Figure 5.14

Of course, given a sentence and a set of *sentominos*, there can be multiple possible representations, corresponding to multiple possible maximal cliques on the graph.

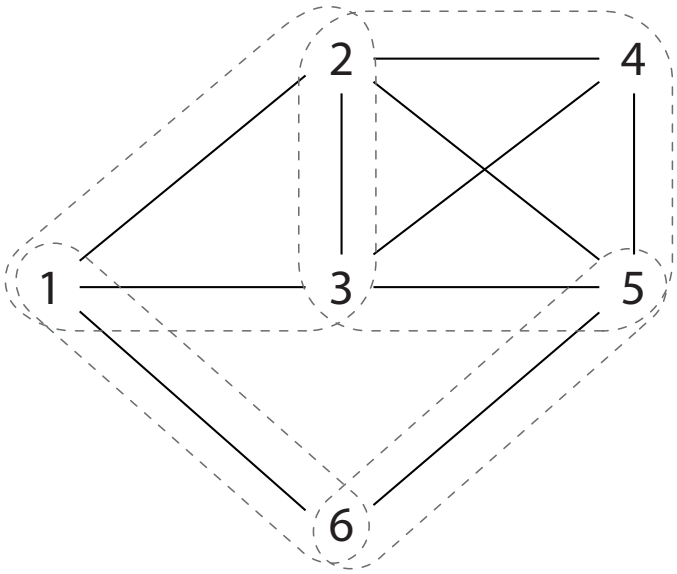


Figure 5.14: The figure shows a simple undirected graph with all the maximal cliques that can be identified in it.

We additionally note that clique detection is considered to be a NP-complete problem: there exists therefore no trivial algorithm to solve it without recurring to brute-force, which is of course not feasible because of exponential growth.

5.3.1.3 *Score representation*

Given our setting, the possible parameters we can use to decide which representation best describes the sentence s are the following:

COVERAGE portion of the sentence covered by the translation

SCHEMATICITY abstractness of the elements occurring in the translation (i.e., lexical items, PoS categories or syntactic relations)

SIZE OF CLIQUE number of constructions concurring to the translation

These three parameters capture different aspects of the translation. More specifically, **coverage** can be regarded as a proxy for the portion of the sentence that is *interpretable* by the listener. **Schematicity**, on the other hand, balances the amount of fine-grained or coarse-grained meaning activated: as it is true that all constructions bear meaning, it is also generally true that the more schematic the patterns, the more coarse-grained their meaning is. Lastly, the *size of clique* can be seen as a very high-level proxy for short-term memory factors.

5.3.2 *Looking through core and periphery*

In order to explore our hypothesis, we sample a set of sentences from the three considered subcorpora. More specifically, we consider 13010 distinct sentences from CHILDES, 12414 from OpenSubtitles, and 12727 from Simple Wikipedia, for a total of total of 37242 distinct sentences.

We then consider *core* and *periphery* sets of constructions extracted for population Π_I, as described in the section above.

For each sentence, we build two graphs of *sentominos*, one using core constructions and the other one using periphery constructions.

5.3.2.1 *Sentominos graphs and sentence complexity*

The number of constructions, and hence the number of *sentominos*, activated for each sentence depends on various factors, including but not limited to the length of the sentence itself.

The size of the resulting graphs is therefore similarly variable. In Figure 5.15, the number of graphs (i.e., sentences) existing for each number of nodes is shown.

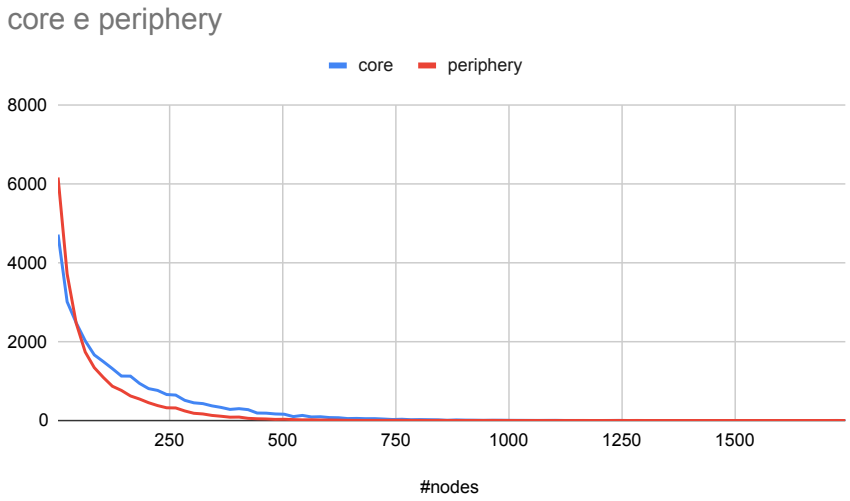


Figure 5.15: On the y axis, the number of sentences (or graphs) existing for the *core* class and (in blue) *periphery* class (in red). The x axis represents the number of nodes in each graph.

As clique-finding is a NP-complete process, we could not compute the maximal cliques for all resulting graphs. We therefore (for computing time reasons) binned the graphs based on their number of nodes and computed maximal cliques for graphs containing up to 141 nodes. The number of graphs for these bins is reported in Table 5.10

We then compare, for each considered bin, the average number of nodes, edges and computing times required for finding all the maximal cliques in the graph: as expected from the NP-completeness

#nodes	core	periphery
2 to 21	4710	6156
22 to 41	3008	3713
42 to 61	2476	2465
62 to 81	2018	1735
82 to 101	1662	1339
102 to 121	1492	1091

Table 5.10: Number of graphs built through *core* and *periphery* subsets.

nature of the problem, both the number of edges and the computing time grow exponentially when increasing the number of nodes in the graph (Table 5.11).

Lastly, we note that the number of nodes does not correlate with the number of edges (Spearman correlation in Table 5.12): the fact that their relation is not obvious confirms that the graphs and their internal structure can be potentially useful representations of sentence complexity and of meaning stratification within the sentence.

5.3.2.2 *Core, periphery and sentence meaning*

As described in Section 5.3.1.3, we computed the following scores for each obtained maximal clique:

COVERAGE number of valorized positions divided by the length of the sentence

SCHEMATICITY we assigned different scores depending on the category of the value for each position of the unified sentence. More specifically, empty positions were assigned 0, lexical items 1/3, part of speech categories 2/3 and syntactic relations 1. The sum of scores was then divided by the length of the sentence

SIZE OF CLIQUE given all the possible representations for a given sentence, we computed the minimum and maximum number of constructions involved. Each representation was then assigned a score as follows:

$$1 - \frac{\text{size of repr} - \min \text{size} + 1}{\max \text{size} - \min \text{size} + 1} \tag{5.13}$$

The "1−" was added to have all scores on the same scale (i.e., the higher the better)

By means of these scores, we computed an aggregated model: we averaged *coverage*, *schematicity* and *size of clique* scores to get an overall weight assigned to the translation.

nodes		core			periphery		
from	to	avg nodes	avg edges	avg computing time	avg nodes	avg edges	avg computing time
2	21	12,420	14,270	0,002	11,325	12,276	0,002
22	41	30,450	102,864	0,006	30,536	100,162	0,005
42	61	50,866	301,025	0,021	51,186	328,426	0,018
62	81	70,891	671,560	0,138	71,159	728,260	0,096
82	101	91,321	1220,370	2,313	90,878	1297,051	0,473
102	121	111,316	1808,597	2,487	111,303	2092,986	1,631

Table 5.11: Average number of nodes, edges and average computing time for finding all maximal cliques in the considered graphs.

nodes			
from	to	core	periphery
2	21	0,777	0,818
22	41	0,629	0,696
42	61	0,389	0,536
62	81	0,308	0,38
82	101	0,267	0,351
102	121	0,18	0,324

Table 5.12: Spearman correlation between number of nodes and number of edges

For each sentence, we scored all its possible translations. This process does not necessarily yield a single best result, as it is possible that different translations are assigned the exact same score. In such cases, we chose randomly among the *best translations* in order to select one.

Our hypothesis was that, given a set of sentences, and selecting a single best translation for each of the sentences in both cases (namely, using core constructions vs. using periphery constructions), we would end up with a smaller set of translations in the case of core constructions than in the case of periphery ones. In other words, more sentences would end up being represented by means of the same translation when using core constructions than in the other case.

As described in Section 5.2.3, we identified 3277 *core* catenae and 6720 *peripheral* ones. Because of this numerical difference, to make the comparison more fair, we restrict to sentences that present both a *core* and a *periphery translation*. This ensures that, both for core and periphery, the graph of that sentence contains less that 122 nodes. In other words, for each of these sentences (9579 in total) we have both a translation with core constructions and a translation with periphery constructions.

Our hypothesis seems to go in the right direction, as we count 7119 unique core translations and 7817 unique periphery translations.

In Table 5.13, we report some examples of the different emerging best translations.

5.4 DISCUSSION

In this chapter, we introduced a representation for language as an abstraction over the linguistic productions of a community of speakers. A key concept when defining language from the usage-based perspective is in fact the variation that exists among individuals and their ability to align on communicative intentions. With respect to this latter aspect, our experiment was of course preliminary as our *artificial speakers* did not interact in any way.

sentence	core	periphery
i will get a job .	@nsubj @aux @root @det @obj _	@nsubj @aux get @det @obj _
i was thinking italian .	@nsubj @aux @root @xcomp _	_PRON @aux @root _ADJ _
he 's a jolly good fellow	@nsubj @cop @det _ @amod @root	_PRON @cop a @advmod @amod @root
They divorced a short time later .	@nsubj @root @det @amod _NOUN _ _	_ @root _DET _ time @advmod _
nicodemus is cooperating beautifully ...	@nsubj @aux @root _ _	_NOUN _ @root @advmod _
A German glassblower invented marble scissors , a device for making marbles , in 1846 .	_ _ @nsubj @root @amod @obj _ @det @obj @mark @acl @obj _ _ _ _ _	@det @amod @nsubj @root _ @obj _ a _NOUN @mark @acl _NOUN _ _ _ _

Table 5.13: Example of translations

We note, however, how we managed to shift from the mainstream view of a Neural Language Model as an average individual speaker by representing each trained instance as a separate individual. This allowed for some structural patterns to emerge.

In particular, we explored two different aspects.

Firstly, we identified a set of shared constructions, both in the homogeneous and inhomogeneous setting. This showed how the amount of *linguistic knowledge* in Neural Language Models that can be considered to be shared across domains accounts only for a share of the entirety of their productions. Moreover, we showed how frequency strongly predicts that a construction will be shared across all speakers, but there are outliers as well. These, in particular low-frequency core constructions, seem to be lexicalized (as opposed to *rather schematic*) patterns, similar to MultiWord Expressions.

Then, we developed a *translation* process that mimics how utterance interpretation might take place across speakers with different constructions. We used the process to highlight differences between *core* and *periphery* constructions and their role in sentence interpretation. With respect to this latter aspect, we have some preliminary indication that using *core* constructions results in fewer distinguishable *meanings* (in our case, sentence structures) than using *periphery* constructions.

The experiments and methods we described here represent just a small step towards a different approach at evaluating Neural Language Models' linguistic productions. Many aspects are left for discussion and further exploration:

- despite having meaning (as in, distributional vectors) in the constructions, we did not explore the distributional space for this first experiment. Different speakers come in fact with different meaning associated to constructions, and as we composed labels (i.e., *sentominos*) to build up a sentence we could as well compose vectors to represent the actual meaning that the speaker intended to convey;

- in particular for the translation process, we have not tested the significance of our results: it remains in fact unclear what would happen when attempting translations based on a completely random set of constructions, for instance. Would the graphs look much different? Would the number of unique translations be significantly higher?

- similarly, we only looked at variation within a population in terms of genre. Exploring different kinds of variation would be a crucial next step to evaluate the usefulness of the procedure. We find two directions to be particularly interesting: exploring variation based on sociolinguistic parameters, and training speakers on artificially created corpora, tailored at testing the ef-

fect of the absence (or the different distribution) of some specific grammatical constructions;

- finally, the created graphs would be interesting to explore per se. The way *sentominos* arrange in the graph represents their linguistic behavior: some might act as hubs, while others might represent bridges.

6 CONCLUSION

Neural Language Models, be it in their vanilla-flavor such as the LSTMs that we implemented in this work, or the more complex Transformer models that are now emerging, are nowadays the state of the art for Language Modelling. This seemed to represent a huge paradigm and methodological shift: surely computational linguistics has always been the result of interplay between theoretical and experimental linguistics, computer science, engineering approaches, cognitive science, philosophy, and many more research fields, but the way the different components of the discipline interact with one another, the balance between the different souls of the subject, has changed during the history of the discipline. We are now probably experiencing one of those turning points, where the contribution of each different component is constantly being redefined and rewritten.

The work we presented in this thesis was born from many thoughts, reflections and discussions on this topic, both inside of the properly defined research *loci* and outside of them, while I slowly developed a personal identity as a computational linguist in my everyday life and activities.

We started from the observation that many latent biases exist in the way Neural Language Models get evaluated, more specifically with respect to their *linguistic competence*. The analysis of literature (Chapter 2) essentially highlighted the following aspects:

- often without explicitly acknowledging it, when evaluating Neural Language Models researchers make use of theoretical categories developed within the Universal Grammar framework: these presuppose a number of assumptions about the properties of the language faculty that do not necessarily apply to the neural computational setting;
- Neural Models show many features that make them suitable for modelling usage-based and cognitively inspired theories;
- there seem to currently be no shared framework to analyse Neural Language Models from a constructionist perspective. A few studies recently emerged in this niche, providing a different perspective on Neural Language Models' linguistic abilities.

For these reasons, we embraced here the Construction Grammars perspective, with the aim of providing new and original points of view to the mainstream evaluation framework. We acknowledge the partiality and limitations of our work, that will be better discussed later in this chapter, but we need to remark here how the lack of a shared methodology also entails the absence or scarcity of tools and resources to rely on. In particular, as Construction Grammar is not very widespread in the computational domain, we embrace the suggestion of Weissweiler, He, et al. (2023) to *adopt a fundamentally different methodology that would establish a standard of evidence/generalisability comparable to GG-based probing*.

Among the reasons why so little has been done in this specific subfield is the lack of data. Construction Grammarians are often concerned with developmental data and small-scale psycholinguistic experiments: the kind of data involved in these settings is very different from what we need to test an automatic and large-scale model. Moreover, to the best of our knowledge, there exists no shared list of constructions or a standardized way of annotating their presence in corpora. Moreover, we feel that the field would benefit from a deeper discussion about the nature of the architectures and models themselves: we discussed the many similarities between usage-based principles and neural approaches to learning, but many points remain unclear, specifically on core aspects and questions. The most recent models are, for instance, necessarily pretrained: what would be the role of this pre-acquired knowledge in a constructionist perspective? And does the task on which architectures are trained matter?

Our main contribution, therefore, was that of introducing CALaMo (Chapter 3): a framework for *looking at* the linguistic abilities of language models in a usage-based perspective. At this point of the discussion, we want to emphasize the verb we used, *'looking at'*, rather than 'evaluating' or 'analysing'. Our constant effort was in fact to rephrase problems in terms of constructionist categories, and this often led to full reconsideration of what could or could not be considered a problem itself, or to discussion about what approach to take on evaluation, what metric would better describe results, or whether a purely qualitatively analysis could be informative enough.

The two exploratory experiments we performed (Chapter 4 and Chapter 5), we believe, showcase these difficulties.

In Chapter 4, for instance, we aimed at tackling two main questions:

(I): what kind of **structures** are abstracted and reproduced by the network, depending on the specific input stream received during training;

(II): how the **abstraction** of *schematic patterns* takes place over time.

Our results essentially confirm the great ability of Neural Models, even in the vanilla-setting we used, to reproduce statistical regularities (question (i)), yet without merely copying their input. This aligns with the constructionist idea that the shape of the input has a causal role in the formation of linguistic categories and structures acquired by speakers, and therefore makes Neural Models natural test benches for usage-based scenarios. As far as question (ii) is concerned, our main contribution was showing how there exists a relation between meaning (i.e., distributional meaning) in the input provided to the Neural Model and the distribution of grammatical patterns during training. The results are preliminary, but open the way to new and interesting ways of tracing the evolution of linguistic abilities of artificial learners.

Chapter 5, instead, introduced the idea of a Neural Language Model as an individual speaker rather than as an idealized average. We therefore tried to work with populations of Neural Models rather than single instances, and dealt with two main questions:

(I) we first set out to confirm the existence of a set of *core* constructions that are learned by all speakers of a community independently of the input they receive and in a statistically significant form;

(II) we then asked how speakers who have acquired different grammars can actually communicate: provided that a *core* set of constructions has been identified.

For our analysis, we considered two different settings (a population of speakers trained on homogeneous data and a population of speakers trained on inhomogeneous data) and focused on describing what could be observed at the group level. For question (i), we provided a report of similarities and differences between the two different settings and tentatively defined *language* at the population level as the set of constructions that are mutually intelligible (i.e., shared) by the whole group. For question (ii), we introduced a representation function that allowed for *transforming* sentences in order to see their linguistic skeleton: performing this operation with shared linguistic knowledge provides an idea of what can be considered common ground for the speakers and what instead constitutes *novel* content.

6.1 LIMITATIONS

One could identify many limitations in our approach, and rightly so.

The main one is certainly due to the choice of relying on a statistical parser for the catenae extraction process. Catenae (and therefore constructions) are in fact identified on the basis of a syntactic dependency tree: this means that all sentences need to be syntactically annotated in order to build a construction. The parser can be considered a linguistic

model itself, and as such it comes with its own biases, due to the kind of data it was trained on. The corpora that we used in this work (child-directed data and network-produced *babblings*) are likely to be very different from the kind of data that the parser was built on. Error rates of statistical parsers are typically higher on out-of-distribution data: this means that our treebanks probably contain a relevant number of sentences for which the parser provides an unreliable representation.

The existence of these cases produces a snowball effect on the whole process, as we end up *seeing* constructions *through the parser's eyes*.

While this currently constitutes the main drawback of our approach, we also remark here that one of the hypotheses on which our model relies is that of having the *linguist-observer* as one of the active agents in the process. As the parser is unable to make sense of ill-formed, children-produced utterances, so it is often the case in real-life scenario. It is, in fact, often hard to apply standard linguistic categories to data with a very peculiar distribution such as child-directed data.

6.2 FUTURE PATHS

As we mentioned multiple times already, our work is in many ways exploratory and the main contribution remains methodological.

The two scenarios in which CALaMo was applied left many further aspects to be explored and questions to be posed.

Out of those scenarios, an interesting future development would be to analyse language acquisition by means of the constructionalization framework provided in Noël (2007) and Traugott and Trousdale (2013). In particular, *constructionalization* refers to the creation of new nodes (i.e., constructions) in the construction. In Noël (2007)'s words, it is defined as *the development through which certain structural patterns acquire their own meanings, so that they add meaning to the lexical elements occurring in them*. Traugott and Trousdale (2013) introduce constructionalization in a framework for diachronic construction grammar as *establishment of a* ***new symbolic association*** *of form and meaning which has been replicated across a network of language users*.

Both these definitions are in line with our approach, with the former one being close to our approach in the first experiment and the latter to our second experiment.

The process identified by Traugott and Trousdale (2013) establishes the existence of modulation effects prior and post constructionalization: Among the pre-constructionalization effects we can cite the loss of compositionality within a construction and the replication of semantic content or syntactic contexts that are connected with the emerging new construction, and the increase in frequencies of these. Once constructionalization has taken place, then, we should observe an expansion of the collocational neighborhood of the construction,

loss of internal analyzability and incorporation into more abstract or schematic nodes.

We believe all these effects could be modeled by means of our distributional analyses.

6.3 TAKE-HOME MESSAGE: NO TECHNOLOGY IS AN ISLAND

Having examined features and limitations of our project, one last question arises and deserves further discussion. We especially thank Prof. Katrin Erk for raising this point during the many meetings that we had over the years. The question touches the very essence of our work: what is it that we are ultimately discussing? Is our model aimed at better modelling human acquisition? Or is it about machines, architectures and their ability to reproduce and mimic human patterns? Or again is our framework just concerned with language models and the ways in which it is conceptualized and used in the field?

There is of course no easy answer to this. We could take a positive or negative stance for each of the aforementioned points, and justify how CALaMo is suited to that specific approach. And that was my answer the first time Katrin came up with this point. In light of the entire work, however, I now realize that none of these stances would represent, on my side, an honest answer to the question.

It is often the case that one begins the PhD journey with the aim of finding a specific and interesting research question to deal with, set out an experimental path and iteratively work on it in order to refine hypotheses and find better results. Questions like this — *what is your research about?* — remind us that this is not the only necessary approach to research. While walking the PhD path, in fact one could be lucky enough to discover how much more there is to research — and, despite all the encountered difficulties and hurdles, I consider myself to be among the lucky ones. There is, in fact, a whole lot about putting things into perspective, reframing problems, connecting approaches and actors involved in the process.

So, is CALaMo about *humans*? It has to be, as language is a uniquely human skill and the data we are using are produced by humans in specific contexts and which specific purposes. Our methods and experiments can therefore be seen as a way of exploring human data, its features and biases. At the same time, human speakers were never involved in the process and we even resisted the temptation to give too much of a human-centric explanation to our results (i.e., the constructions identified during the process), as our focus has been on the method rather than on fitting the model to what we were expecting of it.

Is then CALaMo about *machines*? As one might expect at this point, again, it is and it is not. It is, as it provides insights on what small

scale neural networks can pick up from language, in the era of Large and Pretrained Language Models. At the same time we did not put excessive attention in the selection of hyperparameters, we did not explore different families of models, and we did not properly compare the *goodness* of the model itself. And there is no built-in way in CALaMo to do that.

Lastly, is CALaMo about *Language Modelling*? The answer to this last point depends on what one means by Language Modelling. Making a point about Language Modelling as a computational technique would probably result in a circular argument, and lead to no concrete results. When dealing with language, the hard part resides in defining what language *is*, and the boundaries of the object that we are trying to model. In other words, it is about defining what we precisely expect from the machinery that we are building.

CALaMo was named after the Italian word *calamo* (En. *reed pen*), referring to a writing tool built by cutting and shaping a single reed straw: this tool is the ancestor of our modern pens. The point of CALaMo, I believe, can indeed be better understood when thinking about reed pens: in spite of technological advancements, if we set our mind to describing what a *pen* is and how it works, the flow of work, the various parts of the object, the actors involved — all essential parameters — are still in the picture. What differs is really the relations among them: ballpoints pens have made the writing process easier on a more varied set of surfaces, while styluses are nowadays somehow restricted to specific situations and environments. Analogously, the definition of our object of study, *language*, is continuously changing depending on different parameters in the picture — the computational technology, the medium, speakers and listeners — and relations existing among them. None of these can be looked at on its own, and neither can language as a self-standing object. Our ultimate research question — *what CALaMo is about* — pertains to the relations among components and parameters. What we tried to make is an effort to discuss the technology in the most complex picture we could think of. In the case of Computational Linguistics, the picture has to involve the theoretical framework, the assumptions entailed by the specific perspective we are taking, and ultimately the lines we are drawing around the *language* object.

BIBLIOGRAPHY

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg

2017 "Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks," in *International Conference on LearningRepresentations*.

Alishahi, Afra, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad

2020 (eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Online.

Alishahi, Afra, Grzegorz Chrupała, and Tal Linzen

2019 "Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop," *Natural Language Engineering*, 25, 4, pp. 543-557.

Arehalli, Suhas and Tal Linzen

2020 "Neural Language Models Capture Some, But Not All, Agreement Attraction Effects," in *42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020*.

Artstein, Ron and Massimo Poesio

2008 "Inter-coder agreement for computational linguistics," en, *Computational linguistics (Association for Computational Linguistics)*, 34, 4 (Dec. 2008), pp. 555-596.

Bacon, Geoff and Terry Regier

2019 "Does BERT agree? Evaluating knowledge of structure dependence through agreement relations," *arXiv preprint arXiv:1908.09892*.

Bannard, Colin, Elena Lieven, and Michael Tomasello

2009 "Modeling children's early grammatical knowledge," *Proceedings of the National Academy of Sciences*, 106, 41, pp. 17284-17289.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta

2009 "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora," *Language resources and evaluation*, 43, pp. 209-226.

Bates, Elizabeth, Inge Bretherton, and Lynn Sebestyen Snyder

1988 *From first words to grammar: Individual differences and dissociable mechanisms*, Cambridge University Press, Cambridge, vol. 20.

Bee, Helen L., Lawrence F. Van Egeren, Ann Pytkowicz Streissguth, Barry A. Nyman, and Maxine S. Leckie

1969 "Social class differences in maternal teaching strategies and speech patterns." *Developmental Psychology*, 1, 6p1, p. 726.

Bencini, Giulia M. L. and Adele E. Goldberg

2000 "The Contribution of Argument Structure Constructions to Sentence Meaning," *Journal of memory and language*, 43, 4 (Nov. 2000), pp. 640-651.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell

2021 "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610-623.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael

Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang

2021 "On the Opportunities and Risks of Foundation Models," *ArXiv*, https://crfm.stanford.edu/assets/report.pdf.

Boyd, Jeremy K. and Adele E. Goldberg

2009 "Input Effects Within a Constructionist Framework," *The Modern Language Journal*, 93, 3 (Sept. 2009), pp. 418-429.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei

2020 "Language models are few-shot learners," *Advances in neural information processing systems*, 33, pp. 1877-1901.

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni

2020 "Profiling-ud: a tool for linguistic profiling of texts," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 7145-7151.

Bybee, Joan L. and Paul J. Hopper

2001 *Frequency and the Emergence of Linguistic Structure*, en, John Benjamins Publishing.

Carstensen, Alexandra and Michael C. Frank

2021 "Do graded representations support abstract thought?" *Current Opinion in Behavioral Sciences*, 37 (Feb. 2021), pp. 90-97.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson

2013 "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling" (Dec. 2013), arXiv: 1312.3005 [cs.CL].

Chomksy, Noam

1959 "Review of skinner's Verbal Behaviour," *Language*, 35, pp. 26-58.

Chomsky, Noam

1965 *Aspects of the Theory of Syntax*, 50th ed., The MIT Press, ISBN: 9780262527408.

1968 *Language and Mind*, New York: Harcourt Brace Jovanovich.

Chomsky, Noam

1975 *The logical structure of linguistic theory*, Springer.

1986 *Knowledge of language: Its nature, origin, and use*, Greenwood Publishing Group.

1992 "From Language and Problems of Knowledge," *The Philosophy of Mind: Classical Problems/contemporary Issues*, p. 47.

Chowdhury, Shammur Absar and Roberto Zamparelli

2018 "RNN simulations of grammaticality judgments on long-distance dependencies," in *Proceedings of the 27th international conference on computational linguistics*, pp. 133-144.

Christiansen, Morten H.

2019 "Implicit statistical learning: A tale of two literatures," *Topics in Cognitive Science*, 11, 3, pp. 468-481.

Christiansen, Morten H. and Nick Chater

2016a *Creating language: Integrating evolution, acquisition, and processing*, MIT Press.

2016b "The Now-or-Never bottleneck: A fundamental constraint on language," en, *The Behavioral and brain sciences*, 39 (Jan. 2016), e62.

2016c "The now-or-never bottleneck: A fundamental constraint on language," *Behavioral and brain sciences*, 39, e62.

Clark, Eve V.

2009 *First language acquisition*, Cambridge University Press.

Clark, Herbert H.

1997 "Communal lexicons," in *Context in Language Learning and Language Understanding*, ed. by K. Malmkjær and J. Williams, Cambridge University Press, Cambridge, pp. 63-87.

Cornish, Hannah, Rick Dale, Simon Kirby, and Morten H. Christiansen

2017 "Sequence memory constraints give rise to language-like structure through iterated learning," *PloS one*, 12, 1, e0168532.

Crain, Stephen, Loes Koring, and Rosalind Thornton

2017 "Language acquisition from a biolinguistic perspective," *Neuroscience & Biobehavioral Reviews*, 81, pp. 120-149.

Crain, Stephen and Paul Pietroski

2001 "Nature, nurture and universal grammar," *Linguistics and philosophy*, 24, 2, pp. 139-186.

Crain, Stephen, Rosalind Thornton, and Keiko Murasugi

2009 "Capturing the Evasive Passive," *Language acquisition*, 16, 2 (Mar. 2009), pp. 123-133.

Cristia, Alejandrina, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz

2019 "Child-Directed speech is infrequent in a forager-farmer population: a time allocation study," *Child development*, 90, 3, pp. 759-773.

Croft, William

2001 *Radical Construction Grammar: Syntactic Theory in Typological Perspective*, en, Oxford University Press.

Croft, William and Alan D. Cruse

2004 *Cognitive Linguistics*, en, Cambridge University Press.

Culbertson, Jennifer, Julie Franck, Guillaume Braquet, Magda Barrera Navarro, and Inbal Arnon

2020 "A learning bias for word order harmony: Evidence from speakers of non-harmonic languages," en, *Cognition*, 204 (Nov. 2020), p. 104392.

Culbertson, Jennifer, Paul Smolensky, and Géraldine Legendre

2012 "Learning biases predict a word order universal," en, *Cognition*, 122, 3 (Mar. 2012), pp. 306-329.

Dabrowska, Ewa

2015 "What exactly is Universal Grammar, and has anyone seen it?" en, *Frontiers in psychology*, 6, June, p. 852.

Dąbrowska, Ewa

2004 *Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar*, Edinburgh University Press, Edinburgh, ISBN: 9780748614745, http://www.jstor.org/stable/10.3366/j.ctvxcrgdw.

2013 "Functional constraints, usage, and mental grammars: A study of speakers' intuitions about questions with long-distance dependencies," *Cognitive Linguistics*, 24, 4, pp. 633-665.

Davis, Forrest and Marten van Schijndel

2020a "Discourse structure interacts with reference but not syntax in neural language models," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, pp. 396-407.

2020b "Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 1979-1990, DOI: 10.18653/v1/2020.acl-main.179, https://www.aclweb.org/anthology/2020.acl-main.179.

Deese, James

1966 *The structure of associations in language and thought*, Johns Hopkins University Press, Baltimore.

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning

2014 "Universal Stanford dependencies: A cross-linguistic typology," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 4585-4592, http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf.

De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning

2006 "Generating Typed Dependency Parses from Phrase Structure Parses," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), Genoa, Italy, http://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf.

De Marneffe, Marie-Catherine and Christopher D. Manning

2008 "The Stanford Typed Dependencies Representation," in *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, Coling 2008 Organizing Committee, Manchester, UK, pp. 1-8, https://aclanthology.org/W08-1301.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova

2019 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, pp. 4171-4186.

Dinu, Georgiana, Nghia The Pham, and Marco Baroni

2013 "DISSECT - DIStributional SEmantics Composition Toolkit," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 31-36, https://www.aclweb.org/anthology/P13-4006.

Dittmar, Miriam, Kirsten Abbot-Smith, Elena Lieven, and Michael Tomasello

2008 "Young German children's early syntactic competence: a preferential looking study," en, *Developmental science*, 11, 4 (July 2008), pp. 575-582.

Dunn, Jonathan

2017 "Learnability and falsifiability of Construction Grammars," in *Proceedings of the Linguistic Society of America*, vol. 2, pp. 1-15.

Dupoux, Emmanuel

2018 "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," en, *Cognition*, 173 (Apr. 2018), pp. 43-59.

Dyer, Chris, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith

2016 "Recurrent Neural Network Grammars," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 199-209.

Eisenbeiß, Sonja

2009 47, 2, pp. 273-310, DOI: doi:10.1515/LING.2009.011, https://doi.org/10.1515/LING.2009.011.

Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg

2020 "When Bert Forgets How To POS: Amnesic Probing of Linguistic Properties and MLM Predictions," *arXiv preprint arXiv:2006.00995*.

Elman

2001 "Connectionism and language acquisition," *Language development: The essential readings*.

Erk, Katrin

2012 "Vector space models of word meaning and phrase meaning: A survey," *Language and Linguistics Compass*, 6, 10, pp. 635-653.

Estes, Katharine Graf, Julia L. Evans, Martha W. Alibali, and Jenny R. Saffran

2007 "Can infants map meaning to newly segmented words? Statistical segmentation and word learning," *Psychological science*, 18, 3, pp. 254-260.

Farmer, Thomas A., Jennifer B. Misyak, and Morten H. Christiansen

2012 "Individual differences in sentence processing," *Cambridge handbook of psycholinguistics*, 353, p. 364.

Fazekas, Judit, Andrew Jessop, Julian Pine, and Caroline Rowland

2020 "Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition," *Royal Society Open Science*, 7, 11, p. 180877.

Ferguson, Brock and Casey Lew-Williams

2016 "Communicative signals support abstract rule learning by 7-month-old infants," en, *Scientific reports*, 6 (May 2016), p. 25434.

Fillmore

1976 "Frame semantics and the nature of language," *Annals of the New York Academy of Sciences*.

Fillmore, Charles J.

1988 "The mechanisms of "construction grammar"," in *Annual Meeting of the Berkeley Linguistics Society*, vol. 14, pp. 35-55.

Fousheea, Ruthe, Dan Byrnea, Raquel G. Alhamac, Allyson Ettingerb, Afra Alishahic, and Susan Goldin-Meadowa

submitted "Tracking the onset of productive determiner+ noun combinations in English-learners."

Freudenthal, Daniel, Julian M. Pine, Gary Jones, and Fernand Gobet

2015 "Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions," *Cognition*, 143, pp. 61-76.

Frost, Rebecca L. A., Andrew Jessop, Samantha Durrant, Michelle S. Peter, Amy Bidgood, Julian M. Pine, Caroline F. Rowland, and Padraic Monaghan

2020 "Non-adjacent dependency learning in infancy, and its link to language development," en, *Cognitive psychology*, 120 (Aug. 2020), p. 101291.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy

2019 "Neural language models as psycholinguistic subjects: Representations of syntactic state," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 32-42, DOI: 10.18653/v1/N19-1004, https://aclanthology.org/N19-1004.

Geiger, Atticus, Alexandra Carstensen, Michael C Frank, and Christopher Potts

2022 "Relational reasoning and generalization using nonsymbolic neural networks." *Psychological Review*.

Gergely, György, Harold Bekkering, and Ildikó Király

2002 "Rational imitation in preverbal infants," *Nature*, 415, 6873, pp. 755-755.

Gerken, Louann

2006 "Decisions, decisions: infant language learning when multiple generalizations are possible," en, *Cognition*, 98, 3 (Jan. 2006), B67-74.

2010 "Infants use rational decision criteria for choosing among models of their input," en, *Cognition*, 115, 2 (May 2010), pp. 362-366.

Gertner, Yael, Cynthia Fisher, and Julie Eisengart

2006 "Learning words and rules: abstract knowledge of word order in early sentence comprehension," en, *Psychological science*, 17, 8 (Aug. 2006), pp. 684-691.

Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema

2018 "Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 240-248.

Goldberg, Adele E.

1995 *Constructions: A construction grammar approach to argument structure*, University of Chicago Press.

2006 *Constructions at work: The nature of generalization in language*, Oxford University Press on Demand.

2013 "Constructionist Approaches," in *The Oxford Handbook of Construction Grammar*, ed. by T. Hoffmann and G. Trousdale, Oxford University Press, Oxford, pp. 15-31, DOI: http://dx.doi.org/10.1093/oxfordhb/9780195396683.013.0002.

2015 "Compositionality," in *The Routledge handbook of semantics*, Routledge, pp. 419-433.

2019 *Explain me this: Creativity, competition, and the partial productivity of constructions*, Princeton University Press.

Gomez, Rebecca L.

2002 "Variability and detection of invariant structure," *Psychological Science*, 13, 5, pp. 431-436.

Gomez, Rebecca L. and LouAnn Gerken

1999 "Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge," *Cognition*, 70, 2, pp. 109-135.

Gómez, Rebecca and Jessica Maye

2005 "The developmental trajectory of nonadjacent dependency learning," *Infancy*, 7, 2, pp. 183-206.

Gómez, Rebecca L. and LouAnn Gerken

2000 "Infant artificial language learning and language acquisition," *Trends in cognitive sciences*, 4, 5, pp. 178-186.

Gries, S. T.

2012 "Data in Construction Grammar," *The Oxford Handbook of Construction Grammar*, April 2019, pp. 93-108.

Gulordava, Kristina, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni

2018 "Colorless Green Recurrent Networks Dream Hierarchically," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1195-1205.

Hale, John, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan

2018 "Finding syntax in human encephalography with beam search," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, pp. 2727-2736.

Harris, Zellig S.

1954 "Distributional structure," *Word*, 10, 2-3, pp. 146-162.

Hart, Betty and Todd R. Risley

1995 *Meaningful differences in the everyday experience of young American children.* Paul H. Brookes Publishing.

Hashemzadeh, Maryam, Greta Kaufeld, Martha White, Andrea E. Martin, and Alona Fyshe

2020 *From Language to Language-ish: How Brain-Like is an LSTM's Representation of Nonsensical Language Stimuli?*

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch

2002 "The faculty of language: what is it, who has it, and how did it evolve?" *science*, 298, 5598, pp. 1569-1579.

Hawkins, Robert, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg

2020 "Investigating representations of verb bias in neural language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 4653-4663.

Healy, Alice F. and George A. Miller

1970 "The verb as the main determinant of sentence meaning," *Psychonomic science*, 20, 6 (June 1970), pp. 372-372.

Herbelot, Aurelie

2020 "How to Stop Worrying About Compositionality," *The Gradient*, https://thegradient.pub/how-to-stop-worrying-about-compositionality-2.

Hernandez, Alexia, Sammy Floyd, and Adele E. Goldberg

2019 "Productivity depends on communicative intention and accessibility, not thresholds," *Proceedings of the cognitive science society conference*, 1, pp. 50-57.

Herschensohn, Julia

2009 "Fundamental and gradient differences in language development," *Studies in Second Language Acquisition*, 31, 2, pp. 259-289.

Hilpert, Martin

2014 *Construction grammar and its application to English*, Edinburgh University Press, Edinburgh.

Hochreiter, Sepp and Jürgen Schmidhuber

1997 "Long short-term memory," *Neural computation*, 9, 8, pp. 1735-1780.

Hoffmann, Thomas, Graeme Trousdale, and Ray Jackendoff

2013 *Constructions in the Parallel Architecture*, Oxford University Press, Oxford, ISBN: 9780195396683.

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy

2020 "A Systematic Assessment of Syntactic Generalization in Neural Language Models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1725-1744.

Hudson, Richard A.

1984 *Word grammar*, Blackwell, Oxford.

Hudson Kam, Carla L.

2009 "More than words: Adults learn probabilities over categories and relationships between them," en, *Language learning and development: the official journal of the Society for Language Development*, 5, 2 (Apr. 2009), pp. 115-145.

Hurford, James R.

2000 "Social transmission favours linguistic generalisation," in *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, ed. by C. Knight, M. Studdert-Kennedy, and J. R. Hurford, Cambridge University Press, Cambridge, pp. 324-352.

Jackendoff, Ray

2002 *Foundations of Language: Brain, Meaning, Grammar, Evolution*, OUP Oxford.

Jawahar, Ganesh, Benoıt Sagot, and Djamé Seddah

2019 "What Does BERT Learn about the Structure of Language?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651-3657.

Jolly, Shailza, Sandro Pezzelle, and Moin Nabi

2021 "EaSe: A Diagnostic Tool for VQA Based on Answer Diversity," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2407-2414.

Jolsvai, Hajnal, Stewart M. McCauley, and Morten H. Christiansen

2020 "Meaningfulness Beats Frequency in Multiword Chunk Processing," en, *Cognitive science*, 44, 10 (Oct. 2020), e12885.

Katz, Jerrold J. and Jerry A. Fodor

1963 "The Structure of a Semantic Theory," *Language*, 39, 2, pp. 170-210.

Kay, Paul

1997 *Words and the Grammar of Context*, The Center for the Study of Language and Information Publications, Stanford.

Kay, Paul and Charles J. Fillmore

1999 "Grammatical constructions and linguistic generalizations: the What's X doing Y? construction," *Language*, pp. 1-33.

Klema, V. and A. Laub

1980 "The singular value decomposition: Its computation and some applications," *IEEE Transactions on Automatic Control*, 25, 2, pp. 164-176, DOI: 10.1109/TAC.1980.1102314.

Knowlton, Barbara J. and Larry R. Squire

1994 "The information acquired during artificial grammar learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1, p. 79.

Kuncoro, Adhiguna, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith

2017 "What Do Recurrent Neural Network Grammars Learn About Syntax?" In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1249-1258.

Kuncoro, Adhiguna, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom

2018a "LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, pp. 1426-1436, DOI: 10.18653/v1/P18-1132, https://www.aclweb.org/anthology/P18-1132.

2018b "LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, pp. 1426-1436.

Labov, William

1973 "The boundaries of words and their meanings," *New ways of analyzing variation in English*.

Lakoff, George

1973 *Fuzzy grammar and the performance/competence terminology game*, 2010, De Gruyter Mouton, Berlin, Boston, https://www.degruyter.com/database/COGBIB/entry/cogbib.7026/html.

Lakretz, Yair, Cognitive Neuroimaging Unit, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni

2019 "The Emergence of Number and Syntax Units in LSTM Language Models," in *Proceedings of NAACL-HLT*, pp. 11-20.

Langacker, Ronald W.

1988 "A usage-based model," *Topics in Cognitive Linguistics*.

Lany, Jill and Amber Shoaib

2020 "Individual differences in non-adjacent statistical dependency learning in infants," en, *Journal of child language*, 47, 2 (Mar. 2020), pp. 483-507.

Lau, Jey Han, Alexander Clark, and Shalom Lappin

2017 "Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge," *Cognitive science*, 41, 5, pp. 1202-1241.

Lenci, Alessandro

2008 "Distributional semantics in linguistic and cognitive research," *Italian journal of linguistics*, 20, 1, pp. 1-31.

2018 "Distributional models of word meaning," *Annual review of Linguistics*, 4, pp. 151-171.

Lepori, Michael, Tal Linzen, and R Thomas McCoy

2020 "Representations of syntax [MASK] useful: Effects of constituency and dependency structure in recursive LSTMs," in *Association for Computational Linguistics*.

Lewis, John D. and Jeffrey L. Elman

2001 "Learnability and the statistical structure of language: Poverty of stimulus arguments revisited," in *Proceedings of the 26th annual Boston University conference on language development*, Citeseer, vol. 1, pp. 359-370.

Lewkowicz, David J., Mark A. Schmuckler, and Diane M.J. Mangalindan

2018 "Learning of hierarchical serial patterns emerges in infancy," *Developmental psychobiology*, 60, 3, pp. 243-255.

Li, Bai, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu

2022 "Neural reality of argument structure constructions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7410-7423.

Lidz, Jeffrey and Alexander Williams

2009 20, 1, pp. 177-189, DOI: doi:10.1515/COGL.2009.011, https://doi.org/10.1515/COGL.2009.011.

Lin, Yongjie, Yi Chern Tan, and Robert Frank

2019 "Open Sesame: Getting inside BERT's Linguistic Knowledge," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 241-253.

Linzen, Tal

2020 "How Can We Accelerate Progress Towards Human-like Linguistic Generalization?" In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 5210-5217, https://www.aclweb.org/anthology/2020.acl-main.465.

Linzen, Tal and Marco Baroni

2020 "Syntactic Structure from Deep Learning," en, *arXiv preprint arXiv:2004.10827*, 7, 1 (Jan. 2020), pp. 195-212.

2021 "Syntactic structure from deep learning," *Annual Review of Linguistics*, 7, pp. 195-212.

Linzen, Tal, Grzegorz Chrupała, and Afra Alishahi

2018 (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Brussels, Belgium.

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes

2019 (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Florence, Italy.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg

2016 "Assessing the ability of LSTMs to learn syntax-sensitive dependencies," *Transactions of the Association for Computational Linguistics*, 4, pp. 521-535.

Lison, Pierre and Jörg Tiedemann

2016a "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, pp. 923-929, https://aclanthology.org/L16-1147.

2016b "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles."

MacWhinney, Brian

2000 *The CHILDES Project: Tools for analyzing talk. Third Edition.* Lawrence Erlbaum Associates.

Madabushi, Harish Tayyar, Laurence Romain, Dagmar Divjak, and Petar Milin

2020 "CxGBERT: BERT meets Construction Grammar," in *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics (ICCL), pp. 4020-4032.

Marantz, Alec

2005 22, 2-4, pp. 429-445, DOI: doi:10.1515/tlir.2005.22.2-4.429, https://doi.org/10.1515/tlir.2005.22.2-4.429.

Marchetto, Erika and Luca L. Bonatti

2013 "Words and possible words in early language acquisition," en, *Cognitive psychology*, 67, 3 (Nov. 2013), pp. 130-150.

2015 "Finding words and word structure in artificial speech: the development of infants' sensitivity to morphosyntactic regularities," en, *Journal of child language*, 42, 4 (July 2015), pp. 873-902.

Marcus, G., S. Johnson, K. Fernandes, and J. Slemmer

2004 "Rules, statistics and domain-specificity: Evidence from prelinguistic infants," in *Talk presented at the 29th Annual Meeting of the Boston University Conference on Language Development, Boston, MA*.

Marcus, G. F., S. Vijayan, S. Bandi Rao, and P. M. Vishton

1999 "Rule learning by seven-month-old infants," en, *Science*, 283, 5398 (Jan. 1999), pp. 77-80.

Markman, A. B. and D. Gentner

1993 "Structural Alignment during Similarity Comparisons," *Cognitive psychology*, 25, 4 (Oct. 1993), pp. 431-467.

Marti, Louis, Steven T. Piantadosi, and Celeste Kidd

2019 "Same Words, Same Context, Different Meanings: People are unaware their own concepts are not always shared.," in *CogSci*, pp. 2296-2302.

Marvin, Rebecca and Tal Linzen

2018 "Targeted Syntactic Evaluation of Language Models," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192-1202.

2019 "Targeted Syntactic Evaluation of Language Models," *Proceedings of the Society for Computation in Linguistics (SCiL)*, pp. 373-374.

Masini, Francesca

2016 *Grammatica delle costruzioni*, Carocci, Roma.

Matthews, Danielle and Colin Bannard

2010 "Children's Production of Unfamiliar Word Sequences Is Predicted by Positional Variability and Latent Classes in a Large Sample of Child-Directed Speech," *Cognitive science*, 34, 3, pp. 465-488.

McCauley, Stewart M. and Morten H. Christiansen

2019 "Language learning as language use: A cross-linguistic model of child language development." *Psychological review*, 126, 1, p. 1.

McClelland, James L.

1992 "Can connectionist models discover the structure of natural language," *Minds, Brains and Computers*, pp. 168-189.

McCoy, R. Thomas, Robert Frank, and Tal Linzen

2018 "Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

2020 "Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks," *Transactions of the Association for Computational Linguistics*, 8, pp. 125-140, DOI: 10.1162/tacl_a_00304, https://aclanthology.org/2020.tacl-1.9.

Miller, George A.

1956 "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, 63, 2, p. 81.

Miller, George A. and Walter G. Charles

1991 "Contextual correlates of semantic similarity," *Language and cognitive processes*, 6, 1, pp. 1-28.

Montague

1970 "English as a formal Language," *Linguaggi nella Societa e nella Tecnica*.

Montague, Richard

1970 "Pragmatics and intensional logic," en, *Synthese*, 22, 1-2 (Dec. 1970), pp. 68-94.

Mulder, Kimberley and Jan H. Hulstijn

2011 "Linguistic skills of adult native speakers, as a function of age and level of education," *Applied linguistics*, 32, 5, pp. 475-494.

Newport, Elissa L. and Richard N. Aslin

2004 "Learning at a distance I. Statistical learning of non-adjacent dependencies," *Cognitive psychology*, 48, 2, pp. 127-162.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman

2020 "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection," in *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4034-4043.

Noël, Dirk

2007 "Diachronic construction grammar and grammaticalization theory," *Functions of language*, 14, 2, pp. 177-202.

Nogueira, Fernando

2014– *Bayesian Optimization: Open source constrained global optimization tool for Python*, https://github.com/fmfn/BayesianOptimization.

Nowak, Martin A., Natalia L. Komarova, and Partha Niyogi

2001 "Evolution of universal grammar," *Science*, 291, 5501, pp. 114-118.

O'Grady, William

1998 "The syntax of idioms," *Natural Language & Linguistic Theory*, 16, 2, pp. 279-312.

O'grady, William

2005 *Syntactic carpentry: An emergentist approach to syntax*, Routledge.

Osborne, Timothy J.

2006 "Beyond the constituent-A dependency grammar analysis of chains," *Folia Linguistica*, 39, 3-4, pp. 251-297.

2018 "Tests for constituents: What they really reveal about the nature of syntactic structure," *Language Under Discussion*, 5, 1, pp. 1-41.

Osborne, Timothy J. and Thomas Groß

2012 *Constructions are catenae: Construction grammar meets dependency grammar*.

Osborne, Timothy J., Michael Putnam, and Thomas Groß

2012 "Catenae: Introducing a novel unit of syntactic analysis," *Syntax*, 15, 4, pp. 354-396.

Pannitto, Ludovica

2019–a *CALaMo Toolkit*, https://github.com/ellepannitto/Catenae.

2019–b *LSTM*, https://github.com/ellepannitto/LSTM/tree/master.

Pannitto, Ludovica and Aurélie Herbelot

2020 "Recurrent babbling: evaluating the acquisition of grammar from limited input data," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, pp. 165-176, https://www.aclweb.org/anthology/2020.conll-1.13.

to appear "CALaMo: a Constructionist Assessment of Language Models," in *Proceedings of the 1st Workshop on Construction Grammars and NLP*, Association for Computational Linguistics, https://arxiv.org/abs/2302.03589.

Pannitto, Ludovica and Aurelie Herbelot

2022 "Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition?" *Frontiers in Psychology*, 13, ISSN: 1664-1078, DOI: 10.3389/fpsyg.2022.741321, https://www.frontiersin.org/article/10.3389/fpsyg.2022.741321.

Partee, Barbara H.

2004 *Compositionality in formal semantics: Selected papers*, John Wiley & Sons.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala

2019 "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pp. 8024-8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Perruchet, Pierre and Chantal Pacteau

1990 "Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge?" *Journal of experimental psychology: General*, 119, 3, p. 264.

Perruchet, Pierre, Annie Vinter, Chantal Pacteau, and Jorge Gallego

2002 "The formation of structurally relevant units in artificial grammar learning," *The Quarterly Journal of Experimental Psychology: Section A*, 55, 2, pp. 485-503.

Petrov, Slav, Dipanjan Das, and Ryan McDonald

2012 "A Universal Part-of-Speech Tagset," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, pp. 2089-2096, http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

Pickering, Martin J. and Simon Garrod

2006 "Alignment as the basis for successful communication," *Research on Language and Computation*, 4, pp. 203-228.

2013 "An integrated theory of language production and comprehension," en, *The Behavioral and brain sciences*, 36, 4 (Aug. 2013), pp. 329-347.

Pinker, Steven

1996 *Language learnability and language development*, Harvard University Press, Cambridge, MA.

1999 *Words and rules: The ingredients of language*, Basic Books.

2009 *Language Learnability and Language Development, With New Commentary by the Author: With New Commentary by the Author*, Harvard University Press, vol. 7.

Pustejovsky, James

1991 "The Generative Lexicon," *Computational Linguistics*, 17, 4, pp. 409-441.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever

2019a "Language models are unsupervised multitask learners," *OpenAI blog*, 1, 8, p. 9.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al.

2019b "Language models are unsupervised multitask learners," *OpenAI blog*, 1, 8, p. 9.

Radwańska-Williams, Joanna

2008 "The" native speaker" as a metaphorical construct," in *Metaphors for Learning*, John Benjamins, pp. 139-156.

Rambelli, Giulia, Emmanuele Chersoni, Philippe Blache, Chu-Ren Huang, and Alessandro Lenci

2019 "Distributional Semantics Meets Construction Grammar. towards a Unified Usage-Based Model of Grammar and Meaning," in *Proceedings of the First International Workshop on Designing Meaning Representations*, pp. 110-120.

Ramscar, Michael, Melody Dye, and Stewart M. McCauley

2013 "Error and expectation in language learning: The curious absence of mouses in adult speech," *Language*, 89, 4, pp. 760-793.

Reali, Florencia and Morten H. Christiansen

2005 "Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence," *Cognitive Science*, 29, 6, pp. 1007-1028.

Richards, Brian J.

1990 *Language development and individual differences: A study of auxiliary verb learning.* Cambridge University Press.

Romberg, Alexa R. and Jenny R. Saffran

2010 "Statistical learning and language acquisition," *Wiley Interdisciplinary Reviews: Cognitive Science*, 1, 6, pp. 906-914.

Rubenstein, Herbert and John B. Goodenough

1965 "Contextual correlates of synonymy," *Communications of the ACM*, 8, 10, pp. 627-633.

Sachs, Jacqueline, Barbara Bard, and Marie L. Johnson

1981 "Language learning with restricted input: Case studies of two hearing children of deaf parents," *Applied psycholinguistics*, 2, 1 (Feb. 1981), pp. 33-54.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport

1996 "Statistical learning by 8-month-old infants," *Science*, 274, 5294, pp. 1926-1928.

Saffran, Jenny R., Seth D. Pollak, Rebecca L. Seibel, and Anna Shkolnik

2007 "Dog is a dog is a dog: infant rule learning is not specific to language," en, *Cognition*, 105, 3 (Dec. 2007), pp. 669-680.

Saffran, Jenny R., Janet F. Werker, and Lynne A. Werner

2006 "The infant's auditory world: Hearing, speech, and the beginnings of language," *Handbook of child psychology*.

Salton, Gerard, Anita Wong, and Chung-Shu Yang

1975 "A vector space model for automatic indexing," *Communications of the ACM*, 18, 11, pp. 613-620.

Santolin, Chiara and Jenny R. Saffran

2019 "Non-Linguistic Grammar Learning by 12-Month-Old Infants: Evidence for Constraints on Learning," *Journal of Cognition and Development*, 20, 3, pp. 433-441.

Santoro, Adam, David Raposo, David G. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap

2017 "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems*, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., vol. 30, pp. 4967-4976.

Schütze, Carson

2016 *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*, Language Science Press.

Schütze, Carson T.

2015 *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*, en, Language Science Press.

Schwartz, Dan and Tom Mitchell

2019 "Understanding language-elicited EEG data by predicting it from a fine-tuned language model," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 43-57.

Shen, Yikang, Shawn Tan, Alessandro Sordoni, and Aaron Courville

2018 "Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks," in *International Conference on Learning Representations*.

Silberer, Carina, Sina Zarrieß, Matthijs Westera, and Gemma Boleda

2020 "Humans meet models on object naming: A new dataset and analysis," in *Scott D, Bel N, Zong C, editors. Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 8-13; Barcelona, Spain. Stroudsburg (PA): ACL; 2020. p. 1893-905*, ACL (Association for Computational Linguistics).

Sinclair, Arabella, Jaap Jumelet, Willem Zuidema, and Raquel Fernández

2022 "Structural persistence in language models: Priming as a window into abstract language representations," *Transactions of the Association for Computational Linguistics*, 10, pp. 1031-1050.

Smith, Neil and Nicholas Allott

2016 *Chomsky: Ideas and ideals*, Cambridge University Press.

Snow, Catherine E.

1972 "Mothers' speech to children learning language," *Child development*, pp. 549-565.

Solan, Zach, David Horn, Eytan Ruppin, and Shimon Edelman

2005 "Unsupervised learning of natural languages," *Proceedings of the National Academy of Sciences*, 102, 33, pp. 11629-11634.

Stahl, Aimee E. and Lisa Feigenson

2015 "Observing the unexpected enhances infants' learning and exploration," *Science*.

Straka, Milan and Jana Straková

2018 *Universal Dependencies 2.3 Models for UDPipe (2018-11-15)*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2898.

Street, James A. and Ewa Dąbrowska

2010 "More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers?" *Lingua. International review of general linguistics. Revue internationale de linguistique generale*, 120, 8 (Aug. 2010), pp. 2080-2094.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum

2019 "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 3645-3650, DOI: 10.18653/v1/P19-1355, https://www.aclweb.org/anthology/P19-1355.

Taylor, J. S. H., Fiona J. Duff, Anna M. Woollams, Padraic Monaghan, and Jessie Ricketts

2015 "How Word Meaning Influences Word Reading," *Current directions in psychological science*, 24, 4 (Aug. 2015), pp. 322-328.

Taylor, John R.

2012 *The mental corpus: How language is represented in the mind*, Oxford University Press.

Tiedemann, Jörg

2012 "Parallel Data, Tools and Interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), pp. 2214-2218, http://www.opensubtitles.org/.

Todd, Peyton and Jean Aitchison

1980 "Learning Language the Hard Way," *First language*, 1, 2 (June 1980), pp. 122-140.

Tomasello, Michael

1999 *The Cultural Origins of Human Cognition*, Harvard University Press.

2003 *Constructing a language: A usage-based theory of language acquisition*, Harvard university press.

Tran, Ke M., Arianna Bisazza, and Christof Monz

2018 "The Importance of Being Recurrent for Modeling Hierarchical Structure," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4731-4736.

Traugott, Elizabeth Closs and Graeme Trousdale

2013 *Constructionalization and constructional changes*, Oxford University Press, vol. 6.

Van de Cruys, Tim

2011 "Two multivariate generalizations of pointwise mutual information," in *Proceedings of the Workshop on Distributional Semantics and Compositionality*, Association for Computational Linguistics, pp. 16-20.

Van Deemter, Kees

2010 *Not exactly: In praise of vagueness*, Oxford University Press.

Van Heugten, Marieke and Elizabeth K. Johnson

2010 "Linking infants' distributional learning abilities to natural language acquisition," *Journal of memory and language*, 63, 2, pp. 197-209.

Vitter, Jeffrey S

1985 "Random sampling with a reservoir," *ACM Transactions on Mathematical Software (TOMS)*, 11, 1, pp. 37-57.

Warstadt, Alex and Samuel R Bowman

2020 "Can neural networks acquire a structural bias from raw linguistic data?" In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*, ed. by Stephanie Denison, Michael Mack, Yang Xu, and Blair C Armstrong, cognitivesciencesociety.org.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman

2020 "BLiMP: The Benchmark of Linguistic Minimal Pairs for English," *Transactions of the Association for Computational Linguistics*, 8 (Dec. 2020), pp. 377-392.

Warstadt, Alex, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman

2020 "Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually)," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online.

Wasow, Thomas and Jennifer Arnold

2005 "Intuitions in linguistic argumentation," *Lingua*, 115, 11, pp. 1481-1496.

Wei, Jason, Dan Garrette, Tal Linzen, and Ellie Pavlick

2021 "Frequency Effects on Syntactic Rule Learning in Transformers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 932-948.

Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al.

2022 "Taxonomy of risks posed by language models," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214-229.

Weissweiler, Leonie, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze

2023 "Construction Grammar Provides Unique Insight into Neural Language Models," *arXiv preprint arXiv:2302.02178*.

Weissweiler, Leonie, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze

2022 "The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 10859-10882, https://aclanthology.org/2022.emnlp-main.746.

Wilcox, Ethan, Roger Levy, Takashi Morita, and Richard Futrell

2018 "What do RNN Language Models Learn about Filler–Gap Dependencies?" In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 211-221.

Wittgenstein, Ludwig

1953 *The philosophical investigations*, Blackwell, Oxford.

Yang, Charles and Robert C. Berwick

2017 "Parameter Setting is Feasible," *Linguistic Analysis*, 41, pp. 3-4.

Yang, Charles D.

2004 "Universal Grammar, statistics or both?" *Trends in cognitive sciences*, 8, 10, pp. 451-456.

Yu, Charles, Ryan Sie, Nicolas Tedeschi, and Leon Bergen

2020a "Word Frequency Does Not Predict Grammatical Knowledge in Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 4040-4054, DOI: 10.18653/v1/2020.emnlp-main.331, https://aclanthology.org/2020.emnlp-main.331.

2020b "Word Frequency Does Not Predict Grammatical Knowledge in Language Models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4040-4054.

Yu, Chen and Dana H. Ballard

2007 "A unified model of early word learning: Integrating statistical and social cues," *Neurocomputing*, 70, 13-15, pp. 2149-2165.

Zeman, Daniel

2008 "Reusable Tagset Conversion Using Tagset Drivers," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf.